Abstract

Background

The purpose of the United States Census Bureau is to provide an accurate, detailed, snapshot of the country. That snapshot provides a trove of data that reveals trends about the American people. From that data we can not only conduct research that illuminates the present but create models that predict outcomes. The dataset for this experiment uses fields extracted from the 1994 United States Census for the purpose of determining financial success.

Objective

The goal of this experiment is to take those fields and create a model for predicting financial success. The data in these fields represent both the intrinsic and extrinsic characteristics of an individual. The construction of this model will provide data on which of these categories of characteristics are more influential on an individual's income.

Methods

Before this model can be constructed, the data will need to be prepared to ensure the most relevant outcome. Missing values and outliers will be identified and handled. Categorical data will be discretized with dummy columns created for algorithmic input. Lastly, the predictors will be tested for collinearity to determine if further reduction is possible.

Following preparation, a series of analyses will be conducted to determine individual correlation between predictor variables and the outcome variable. The outcome variable and measure of financial success will be a binary field denoting whether an individual's income is above $50,000 USD. After analyzing potential trends, different combinations of attributes will be tested and selected. The best attributes will be used for the algorithm. Because the output is binary and disqualifies other common models, logistic regression will be implemented for this dataset.

Results

The analysis demonstrated that extrinsic attributes more likely were predictive of financial success. Certain intrinsic characteristics like race and sex were among the highest ranked features. However, these features accounted for only 20% of the total. 80% of the list was composed of extrinsic characteristics such as education, marriage, and occupation.

Conclusion

The disparity between intrinsic and extrinsic predictor variables was highly unexpected. An even split was thought to be the more likely outcome. These results are vindicating to those who believe that the choices made in one's life are the most deterministic of success.

Introduction

Significance

As people of color in higher education in the United States, we are particularly aware of and determined to overcome societal boundaries to personal success. One of the main barometers for that success is financial health. Some of the boundaries that can impede that health are implicit bias, institutional discrimination, and overt inequity. These boundaries are determinant on a person's intrinsic characteristics – race, gender, nationality, etc. We look to measure these immutable attributes against acquired attributes, such as education, to determine their significance in predicting financial success.

The analysis of the attributes in this experiment is not far removed from the purpose of data science: to make our lives better. In this instance, how to live. Should you walk through life believing your choices don't matter? Should you further your education or get married? While this experiment cannot conclusively answer any of these questions, it can shed light on the trends that exist in the US.

Dataset

The dataset is derived from the 1994 United States census database consisting of over 32,000 records. The data consists of variables that pertain to an individual's intrinsic attributes: age, sex, race, and native country. The table below shows a more detailed description of these attributes:

age	17<= Values <=90
sex	Female, Male
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holland-Netherlands

Additionally, the data identifies an individual's extrinsic attributes: marital status, work class, education, occupation, relationship, capital gains, capital losses, and hours worked per week. The table below shows a more detailed description of these attributes:

Marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

education	Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Prof-school, Some-college, Assoc-acdm, Assoc-voc, Bachelors, Masters, Doctorate
education-num	Numerical Representation of education column. Lowest to Highest: 1 - 16
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
capital gains	Profit from an investment or property sale - Values >= 0
capital losses	Loss from an investment or property sale - Values >= 0
hours_per_week	1<= Values <=99

Finally, the dependent variable for the experiment is the income column:

income (less than or greater than 50k)	<=50K, >50K

Methodology

Objective

The goal of this experiment was to identify the values in each column that correlated the most with the >50K value in the income column. Can we use these attributes to predict that result? Furthermore, can a comparison be drawn between intrinsic and extrinsic attributes?

Data Munging

To distill the most relevant result, extraneous and faulty data must be removed. Before coming down to the intrinsic and extrinsic columns listed above, two additional columns (final weight and relationship) were expunged using the 'drop' method. Final weight was a continuous result column from a previous analysis and had no value in this experiment. Without performing collinearity evaluation, relationship clearly represented data that could already be interpreted from marital-status and sex and was therefore redundant.

With the width of the dataset decreases and the volume reduced, there were less variables to muddy the analysis. The 'isnull' method concluded that the dataset contained no missing values. Therefore, there was no need to drop values or use imputation. Next the dataset was evaluated for outliers. By generating the IQR for each continuous data column, we can see the 'spread' of the data. Using these values, we can determine if a column has values significantly out of range that would skew the data. This was visualized with a box-and-whisker plot.

Dimension Reduction

While this dataset was not plagued by datatype variety, the range of values in each column were not convenient for analysis. Continuous data columns like age were reduced into a series of ranges or "buckets". Categorical data columns like work class, education, and native country were categorized into smaller groups for easier digestion. Attributes such as native-country had to be thoughtfully categorized with consideration towards occurrence in the dataset, region, and representation. Multiple iterations were required before reaching a balanced result. Although this dataset only used string and integer values, the chosen algorithm requires purely numerical input. To achieve this, dummy columns were generated for all discrete columns. By reducing the range of values for these columns through discretization, the minimum number of dummies were generated.

Finally, collinearity was evaluated. Just as relationship was dropped for effectively redundant data, the rest of the columns must be compared to demonstrate each attribute's individual significance to the model. Using the 'corr' method, a table was generated to compare each attribute's correlation to every other attribute. While none warranted removal, the inverse correlation between 'income_ >50K' and 'income_ <=50K' would be a problem for the model later. The latter column didn't need to be created for the purpose of the model but that is the weakness of using a bulk generator method like 'get_dummies' to create columns.

Data Visualization

Once the data was prepared, the next step was to explore the data for trends. By implementing the 'Seaborn' module, comparisons between dependent and independent variables were able to be visualized. Plotting the proportion between those who earned above 50K per year and those who didn't by class allows us to make early predictions about the most significant features. Trend interpretations validated some, but subverted other initial assumptions about the data going into the experiment.

Classification

Without knowing the limitations of conducting an experiment with a binary dependent, the first series of trials attempted to use linear regression and Naïve Bayes classification algorithm. After concluding with results ranging from 6% to 8% accuracy, further research was done to find a suitable model. Without a normally distributed or at least continuous dependent variable, the assumptions that create an accurate linear model are violated and thus result in the initial outcome. A further problem was the imbalance in the income >50K column. Using the 'group by' method showed that over 85% of the individuals in the dataset were negative for the outcome variable.

Logistic regression is an algorithm suited to binary dependent variables. Before selecting predictor variables, the binary dependent would need to be balanced in order to produce a more relevant model. The predictor variables are the continuous and dummy variables columns because the model requires numeric input. To achieve a balanced outcome the 'imblearn' module was imported to oversample the outcome variable. The goal was to sample the minority class: income_ >50K with value 0 (equivalent to income < 50K) until the classes are equal. After splitting, training, and fitting the data, the proportion of positive and negative column results were both 50%. The weakness of this technique is that it is artificial and actual frequency of appearance is lost.

Once the data was balanced, Recursive Feature Selection (RFE) was used in conjunction with the Logistic Regression to reduce the pool of inputs to the best predictor variables and construct the most accurate model. RFE reconstructs the logistic regression model with different features recursively using increasingly smaller sets of features until best performers are chosen. An array is produced that ranks each attributes effect on income with the best ranked as 1 and True in the Boolean array. The dummy column 'income_ <=50K' was shown earlier to be negatively correlated with the outcome variable. Removed here to prevent a singular array error. After aggregating these selected features, we can train the data, run the regression, and fit the model. After the construction of the model, the results are tested for accuracy and displayed.

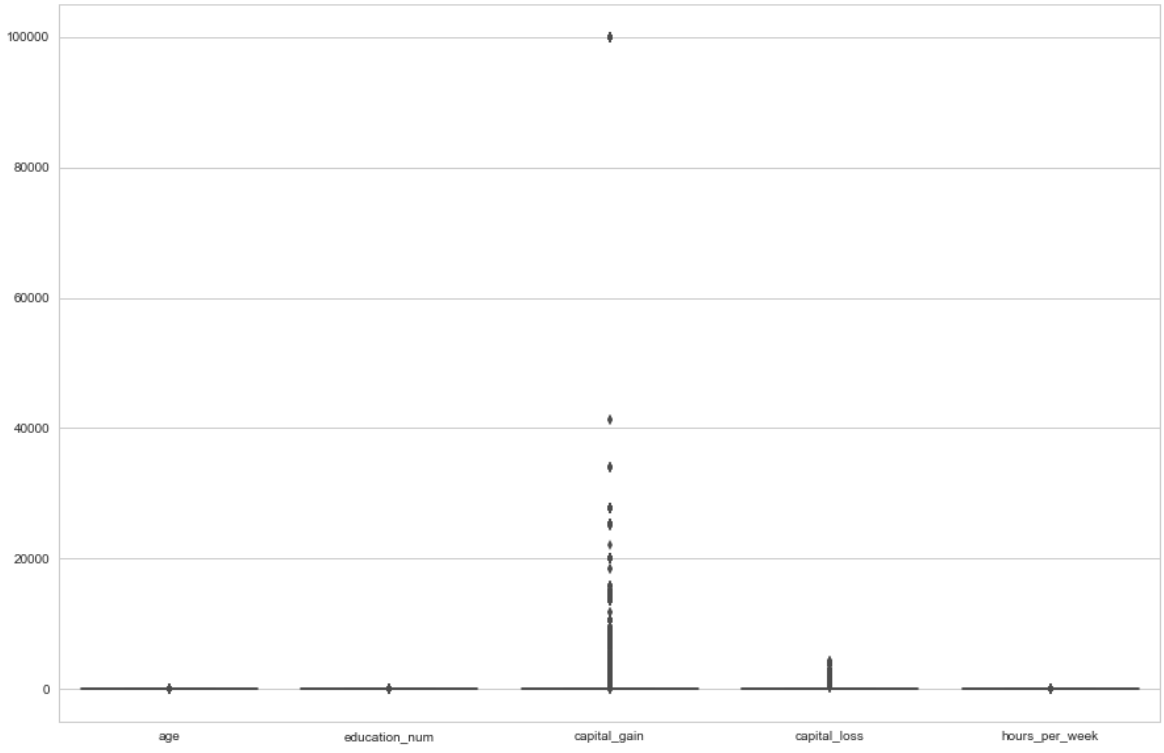
Results

Missing Values

None. No additional steps (imputation, removal) required.

Outliers

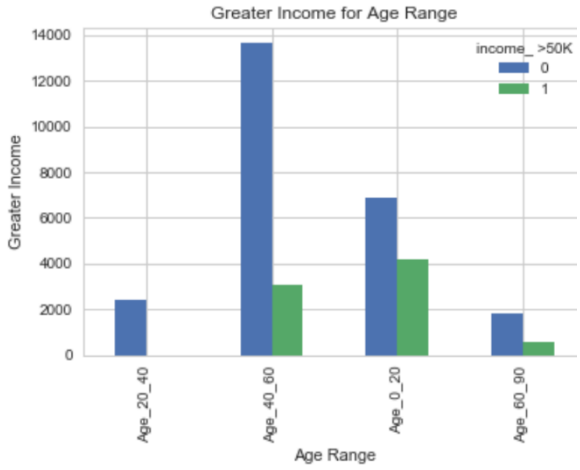
Based on the interquartile range, Age had the widest spread between all the columns. Additionally, Capital-Gains also displayed outlier values (some earning about 100,000). These values were kept because they were still within logical bounds and did not display invalid input. Dropping either would alter the results and assumptions of the experiment.



Feature Performance

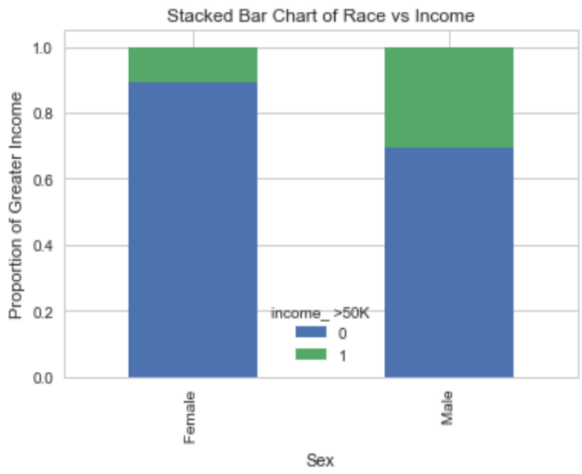
Intrinsic

Age



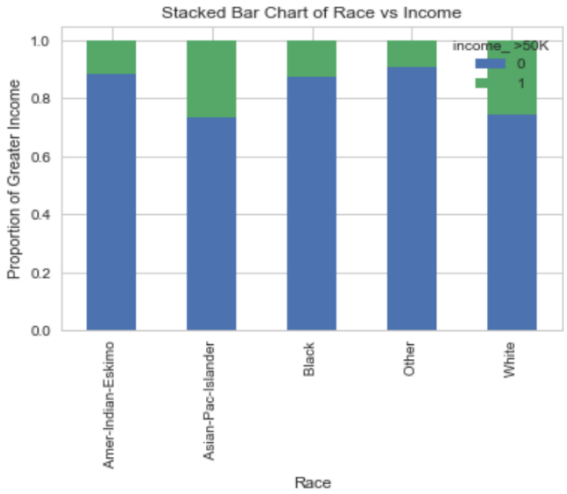
Unexpected outcome. No occurrences of individuals between 20 and 40 earning more than 50K. Possible economic explanation.

Sex



Females have a much lower proportion of earners making above 50K compared to males.

Race



Whites and Asian Pacific Islanders have the highest proportion of earners above 50K.

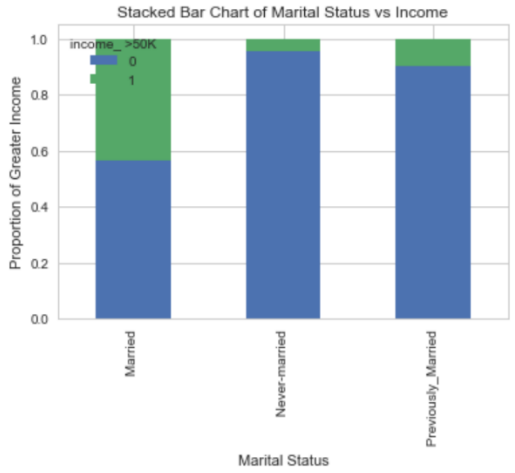
Native Country

Asians (South, Western, East) and Canadians have the highest proportion of >50K earners.

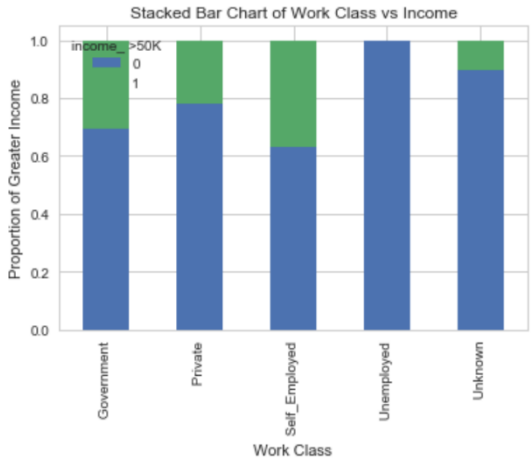
Extrinsic

Marital Status

Work Class

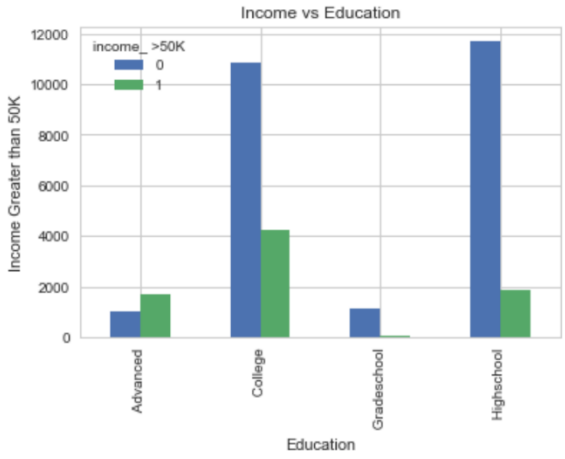


Almost 50% of married individuals are recorded making more than 50K. Unmarried individuals have a drastically lower proportion.



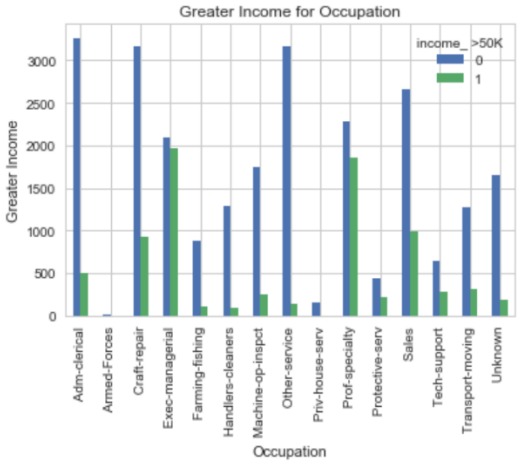
Self-employed individuals have the greatest proportion of >50K earners. Expectedly, the unemployed has no high earners.

Education



An Advanced education is the only factor in this report that has over 50% proportion of earners making above 50K.

Occupation



Executive, Managerial, and Professional Specialty have the high proportion of successful earners.

Feature Selection

Final list of columns (including dummies) selected for predictor variables, as generated by RFE:

```
['education_num', 'workclass_ Unknown', 'education_ Advanced', 'education_
College', 'education_ Gradeschool', 'marital_status_ Married',
'marital_status_ Never-married', 'marital_status_ Previously_Married',
'occupation_ Adm-clerical', 'occupation_ Farming-fishing', 'occupation_
Handlers-cleaners', 'occupation_ Other-service', 'occupation_ Prof-
specialty', 'occupation_ Sales', 'occupation_ Unknown', 'race_ White', 'sex_
Male', 'native_country_ United-States', 'Age_buckets_Age_20_40']
```

Metrics

Logistic Regression

The initial goal was to achieve a model with an accuracy above 80%

```
#Predict test set results
#Calculate Accuracy

y_pred = logreg.predict(X_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_tes
```

```
Accuracy of logistic regression classifier on test set: 0.83
```

Confusion Matrix

```
[[4001 1141]
 [ 643 4603]]
```

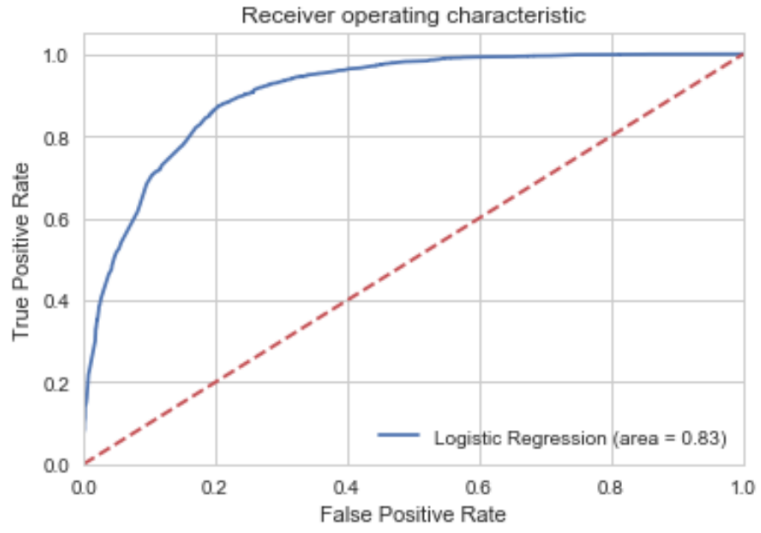
There are **8604** (4001 + 4603) correct predictions and **1784** (643 + 1141) incorrect predictions.

Classification Report

	precision	recall	f1-score	support
0	0.86	0.80	0.83	5142
1	0.81	0.87	0.84	5246
avg / total	0.84	0.83	0.83	10388

Receiver Operating Characteristic (ROC) curve

Used with binary classifiers. Dotted line represents the ROC curve of a purely random classifier (good classifier stays as far away from that line as possible).



Discussion

It is important to note that this report was done from the perspective of three never-married, black, male, students, in the age range of 20-40 in 2018. This lack of diversity may have introduced bias into the preparation of the data and the construction of the model.

From our initial point of view, $50,000 a year was not that much money for an annual salary, nor was it an appropriate amount of money to act as a benchmark for financial success. We later came to understand that $50,000 in 1994 USD is equivalent to approximately 85,000 today. When stated in those terms, our output value seemed like a much more acceptable metric. However, that initial doubt in the veracity of the data might have caused a modification of the experiment or disregarding of the dataset.

Although we decided to keep the dataset, certain caveats must be taken regarding results drawn from it. From our results, we saw a significant disparity between male and female income. Progress made in the last 20+ years, such as the Lilly Ledbetter Fair Pay act of 2009, would presumably limit that disparity.

Being college-aged with aspirations of success, emphasis was placed on age and higher education to be strong determinants of income. Furthermore, we believed that the proportion of individuals making above 50K would scale with age. Surprisingly the data revealed that not only did no individuals in our age group meet this standard, but that there were individuals below our age group who did. However, from the list of selected features, several of the attributes chosen to be the best predictors were categorized under education. Finally, among all attributes, advanced education status was the only category to have a proportion of high earning individuals above 50%.

Other extrinsic attributes such as occupation, marital status, and native country resulted in expected outcomes. Specialized, executive, and white-collar jobs showed the greatest correlation with high income. Marital status was more determinant than presumed, with all marital status' being among the best features. While native country being United States resulted as a selected feature, other regions (specifically in Asia) were proportionally comparable. Much like the income column, these values are highly imbalanced, with over 90% of individuals being native born. While oversampling might have been an option, too much artificial data would have been generated, especially for a single attribute.

Our initial assumptions led us to believe that white males being the most privileged members of American society would result in a high correlation between those attributes and financial success. While the final selected features did include 'race_White' and 'sex_Male' as strong predictors for income greater than 50K, the rest of the results were encouraging. These attributes were ultimately, only a few of the intrinsic attributes that were selected. In fact, only 20% of the list of features were intrinsic. From these results we can infer that the extrinsic attributes, the choices we make in life, are the greatest predictors for financial success.

Contribution

Anthony Harris: Model, Metrics, Visualizations, Abstract, Methodology, Formatting

Samuel Williams: Dimension Reduction, Introduction, Discussion

Emanuel Timbo: Dataset, Data Preparation, Background Research

References

- Census Income Dataset: https://archive.ics.uci.edu/ml/datasets/Census+Income
- Outliers To Drop or not To Drop: https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/
- Why shouldn't I use Linear Regression if my outcome is binary?: http://thestatsgeek.com/2015/01/17/why-shouldnt-i-use-linear-regression-if-my-outcome-is-binary/
- Building a logistic regression in Python Step-by-Step: https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
- Lilly Ledbetter Fair Pay Act: https://en.wikipedia.org/wiki/Lilly_Ledbetter_Fair_Pay_Act_of_2009
- Balancing Data: https://www.ibm.com/developerworks/library/ba-1608balancing-spss-modeler-trs/index.html
- Pandas Documentation: https://pandas.pydata.org/pandas-docs/stable/index.html
- Oversampling and imblearn module: https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html