# Analytics Report on Salary Prediction

**Sean Whalen**

## Executive Summary

This report explores the predictive capability of 14 given variables to determine whether or not an individual within a sample population earns more than 50K a year. We will begin by reviewing the data set for breadth and quality to establish its reliability. Then, we can best interpret the value of the conclusions found between the predictive capability of the given variables as it relates to identifying an individual earning more than 50K a year using IBM SPSS Modeller.

## The Prediction Problem

Many stereotypes or assumptions are made about an individual's success given their demographics whether that be race, sex, level of education, etc. These are, by definition, oversimplified concepts that do not provide a true representation of individual economic sustainability. Using the provided sample data set "Adult.xls", I decided to perform descriptive analytics to best understand the correlation at play between these variables and the target before running a cluster analysis, and developing further models to better identify this set's predictive capability.

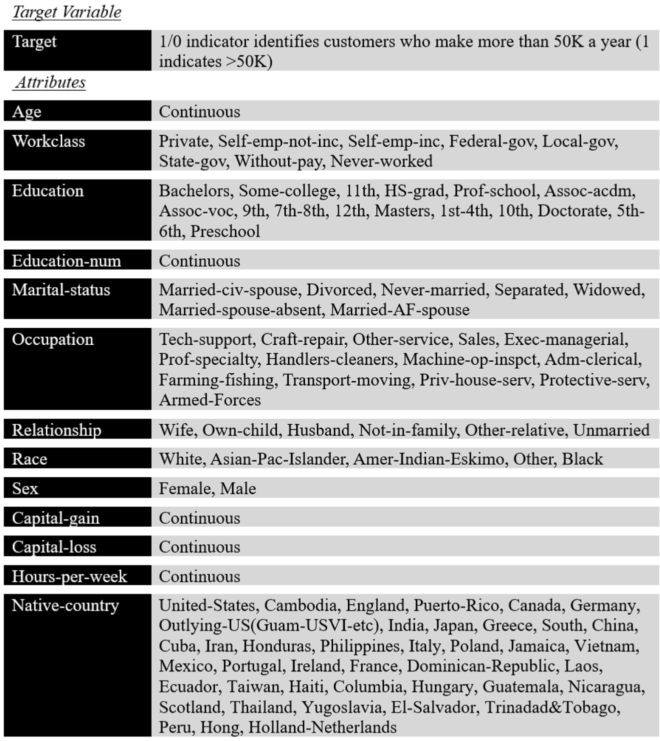
Key relationships to explore in this report will include education level, race, sex, and native country as well as personal relationships, and choice of occupation as they relate to an economically stable salary of 50K or greater.

## The Data Set

This data set came from UCI Machine Learning Repository, a dated, but reliable university repository for databases and data centers used by the Machine Learning community. The sample population in this case contains 32561 individuals broken down across 14 fields with quantitative variables such as Age, Education number, Capital Gain, Capital Loss, and Hours per Week while the rest are qualitative.

The data dictionary pictured above provides all of the variables within the data set and a description of their conditions. The quality was good for most fields with an understanding that the quantitative fields Capital Gain, Capital Loss, Hours per week, and Education Number are unclearly defined. The most significant discrepancy with the data set are the blank values identified by a question mark that are prevalent in the fields workclass, occupation, and native country.

There were 1837 (5.64% of total) question marks in workclass out of the 32561 records, another 1844 (5.66% of total) question marks in the occupation field, and 584 (1.79% of total) in the native country field.  These are not huge numbers but we should note that these values are missing at random without any discernible pattern.  Any outcomes gleaned from analysis using these fields will have degrees of bias and should be measured with that in mind.



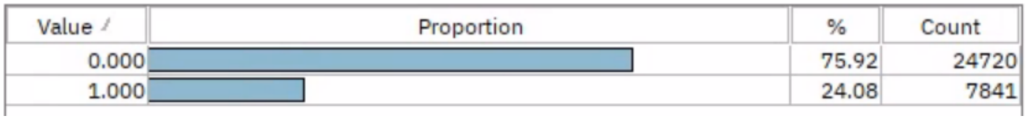
*Target Variable*

| | |
|---|---|
| Target | 1/0 indicator identifies customers who make more than 50K a year (1 indicates >50K) |

*Attributes*

| | |
|---|---|
| Age | Continuous |
| Workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| Education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| Education-num | Continuous |
| Marital-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| 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 |
| Relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| Race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| Sex | Female, Male |
| Capital-gain | Continuous |
| Capital-loss | Continuous |
| Hours-per-week | Continuous |
| 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 |

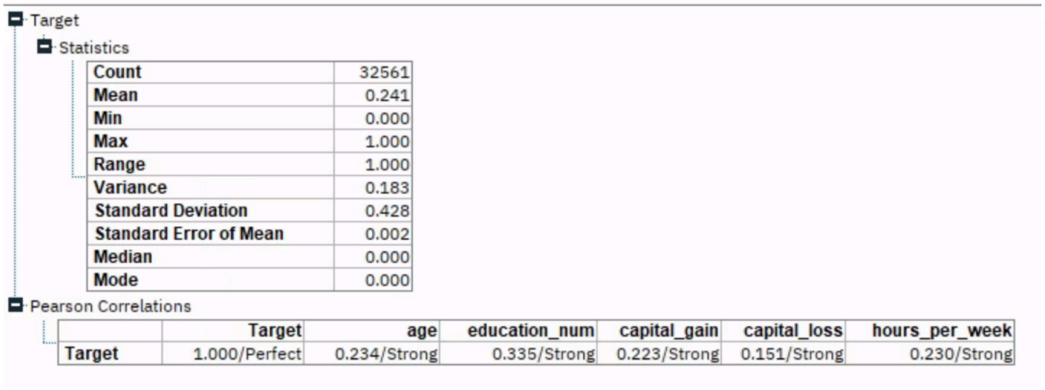
## Baseline for Prediction Goal



| Value | Proportion | % | Count |
|---|---|---|---|
| 0.000 | | 75.92 | 24720 |
| 1.000 | | 24.08 | 7841 |

Based on the distribution chart above, only 24.08% of our sample population totalling 32561 meet the target salary of >50K per year.

The only qualitative variables that can be effectively reviewed are fewer overall, but we can take away from the below that all are strongly, positively correlated with Education num and age being the strongest.



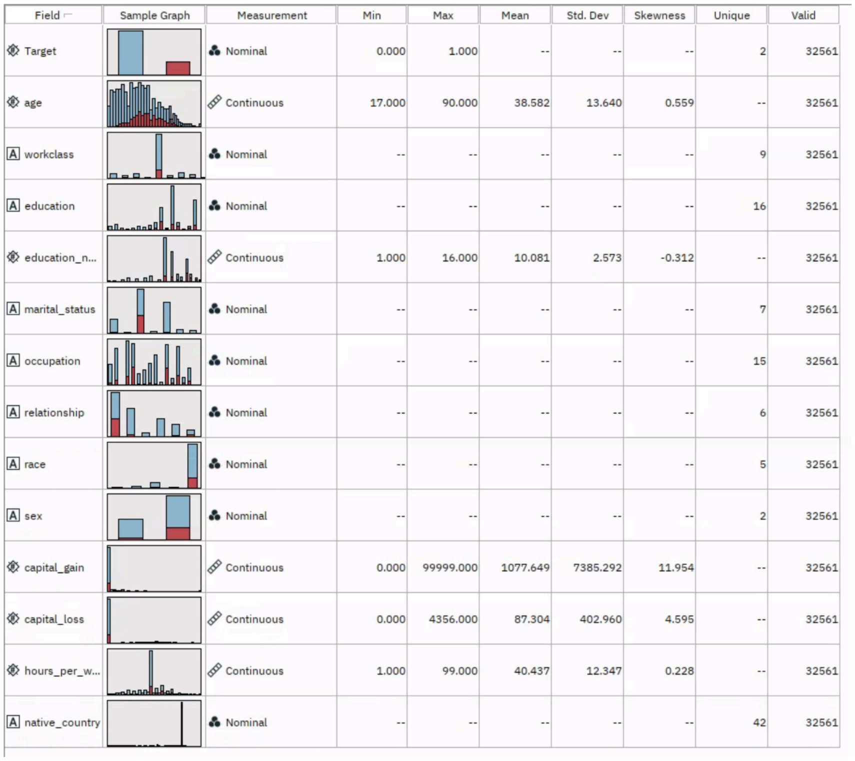
Target

Statistics

| | |
|---|---|
| **Count** | 32561 |
| **Mean** | 0.241 |
| **Min** | 0.000 |
| **Max** | 1.000 |
| **Range** | 1.000 |
| **Variance** | 0.183 |
| **Standard Deviation** | 0.428 |
| **Standard Error of Mean** | 0.002 |
| **Median** | 0.000 |
| **Mode** | 0.000 |

Pearson Correlations

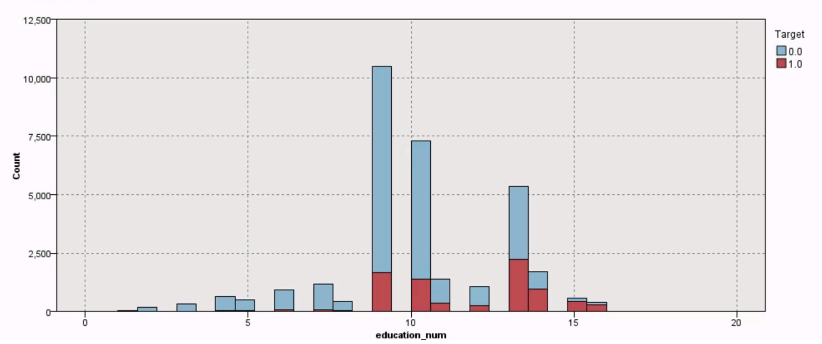
| | Target | age | education_num | capital_gain | capital_loss | hours_per_week |
|---|---|---|---|---|---|---|
| **Target** | 1.000/Perfect | 0.234/Strong | 0.335/Strong | 0.223/Strong | 0.151/Strong | 0.230/Strong |

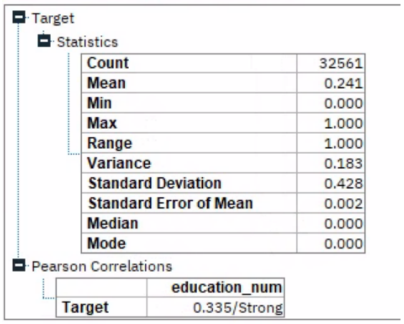


| Field | Sample Graph | Measurement | Min | Max | Mean | Std. Dev | Skewness | Unique | Valid |
|---|---|---|---|---|---|---|---|---|---|
| Target | | Nominal | 0.000 | 1.000 | -- | -- | -- | 2 | 32561 |
| age | | Continuous | 17.000 | 90.000 | 38.582 | 13.640 | 0.559 | -- | 32561 |
| workclass | | Nominal | -- | -- | -- | -- | -- | 9 | 32561 |
| education | | Nominal | -- | -- | -- | -- | -- | 16 | 32561 |
| education_n... | | Continuous | 1.000 | 16.000 | 10.081 | 2.573 | -0.312 | -- | 32561 |
| marital_status | | Nominal | -- | -- | -- | -- | -- | 7 | 32561 |
| occupation | | Nominal | -- | -- | -- | -- | -- | 15 | 32561 |
| relationship | | Nominal | -- | -- | -- | -- | -- | 6 | 32561 |
| race | | Nominal | -- | -- | -- | -- | -- | 5 | 32561 |
| sex | | Nominal | -- | -- | -- | -- | -- | 2 | 32561 |
| capital_gain | | Continuous | 0.000 | 99999.000 | 1077.649 | 7385.292 | 11.954 | -- | 32561 |
| capital_loss | | Continuous | 0.000 | 4356.000 | 87.304 | 402.960 | 4.595 | -- | 32561 |
| hours_per_w... | | Continuous | 1.000 | 99.000 | 40.437 | 12.347 | 0.228 | -- | 32561 |
| native_country | | Nominal | -- | -- | -- | -- | -- | 42 | 32561 |

Most variables through the Data Audit node above displayed a right skewed distribution with native country, sex, race, education number, education, and workclass preferring a left skewed distribution.

Histogram of Education_num overlaid by Target




Target

Statistics

| | |
|---|---|
| Count | 32561 |
| Mean | 0.241 |
| Min | 0.000 |
| Max | 1.000 |
| Range | 1.000 |
| Variance | 0.183 |
| Standard Deviation | 0.428 |
| Standard Error of Mean | 0.002 |
| Median | 0.000 |
| Mode | 0.000 |

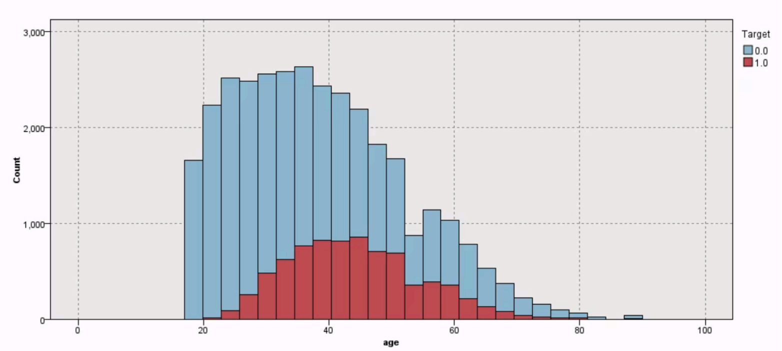
Pearson Correlations

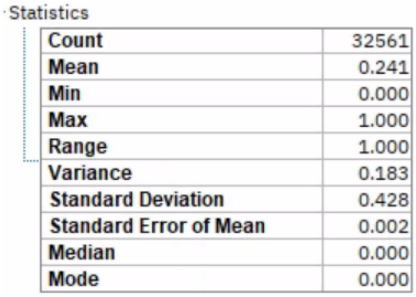
| | education_num |
|---|---|
| Target | 0.335/Strong |

The Education_num field assigns a numerical value that directly correlates to a level of education attainment per individual within the sample population. Despite the Nine indicates 'HS grad' and at that level of education, only 15.95% of the sample population can be accurately predicted to fit our target. However, this predictive capability increases to 19.02% at the 'some college' (10) metric, 26.12% at the 'Associates vocational', falling to 24.84% at the 'Associates degree', before increasing again to 41.48% at the 'bachelors' level and eclipsing 50% predictive accuracy for each level beyond.

The greater the level of education attained within the sample population, the greater the chance that the individual will acquire the skills needed to retain a position that earns >50K a year.

Histogram of Age overlaid by Target




Statistics

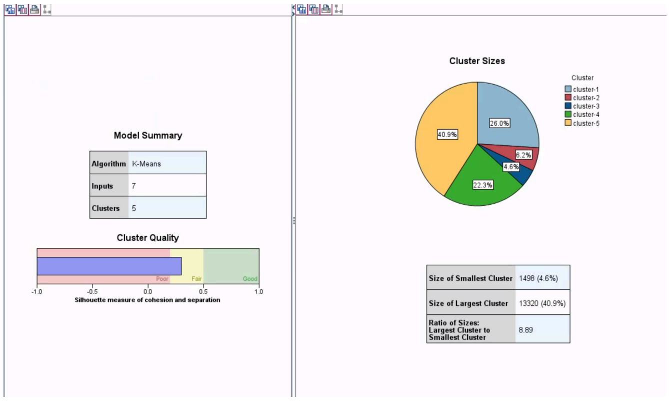
| | |
|---|---|
| Count | 32561 |
| Mean | 0.241 |
| Min | 0.000 |
| Max | 1.000 |
| Range | 1.000 |
| Variance | 0.183 |
| Standard Deviation | 0.428 |
| Standard Error of Mean | 0.002 |
| Median | 0.000 |
| Mode | 0.000 |

The histogram for age proves to be accurate to 24.06% of the sample population aged roughly 33.06 with an increasing level of efficacy, peaking at 41.49% at age 50.58 before leveling out at 24.44% by age 65. Between 33 and 65, we have a >24% chance of predicting our target variable with this data set.

*Part II: Cluster Analysis*

Age, Education Number, Capital Gain, and Hours per week all had fairly strong pearson correlations so I would like to find a way to group them and apply a segmentation. As broken down above, Education Number can cover the education variable since they mirror the same values. I found the Distribution of sex to be convincing in showing that 30.57% of males and only 10.95% of females were identified by the target. A slightly lower 24.58% of the sample population in the US, 25.59% of white, 12.39% of black all identified as targets. The marital status field also seemed to have a fair breakdown.

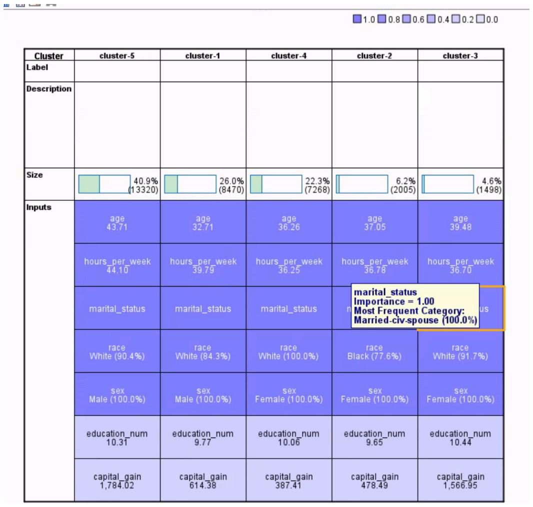
Using these points to start with I created a cluster using age, education number, capital gain, hours per week, sex, race, and marital status.



Model Summary

| | |
|---|---|
| Algorithm | K-Means |
| Inputs | 7 |
| Clusters | 5 |

| | |
|---|---|
| Size of Smallest Cluster | 1498 (4.6%) |
| Size of Largest Cluster | 13320 (40.9%) |
| Ratio of Sizes: Largest Cluster to Smallest Cluster | 8.89 |

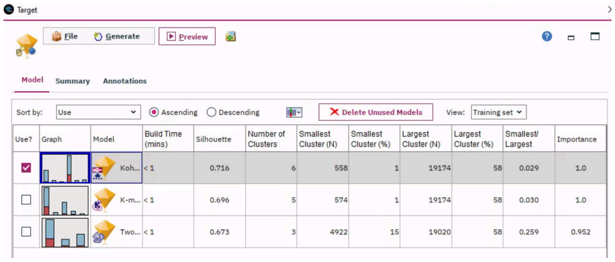


| Cluster | cluster-5 | cluster-1 | cluster-4 | cluster-2 | cluster-3 |
|---|---|---|---|---|---|
| Label | | | | | |
| Description | | | | | |
| Size | 40.9% (13320) | 26.0% (8470) | 22.3% (7268) | 6.2% (2005) | 4.6% (1498) |
| Inputs | age<br>43.71 | age<br>32.71 | age<br>36.26 | age<br>37.05 | age<br>39.48 |
| | hours_per_week<br>44.10 | hours_per_week<br>39.79 | hours_per_week<br>36.25 | hours_per_week<br>36.78 | hours_per_week<br>36.70 |
| | marital_status | marital_status | marital_status | n | us |
| | race<br>White (90.4%) | race<br>White (84.3%) | race<br>White (100.0%) | race<br>Black (77.6%) | race<br>White (91.7%) |
| | sex<br>Male (100.0%) | sex<br>Male (100.0%) | sex<br>Female (100.0%) | sex<br>Female (100.0%) | sex<br>Female (100.0%) |
| | education_num<br>10.31 | education_num<br>9.77 | education_num<br>10.06 | education_num<br>9.65 | education_num<br>10.44 |
| | capital_gain<br>1,784.02 | capital_gain<br>614.38 | capital_gain<br>387.41 | capital_gain<br>478.49 | capital_gain<br>1,566.95 |

marital_status
Importance = 1.00
Most Frequent Category:
Married-civ-spouse (100.0%)

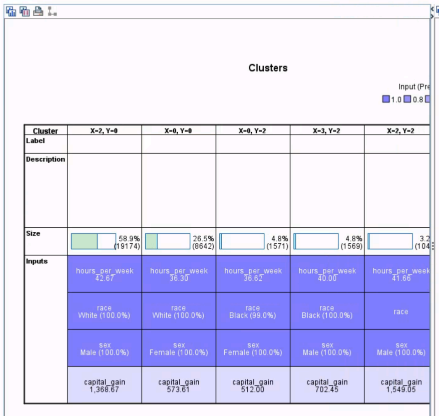
This first attempt proved to be ineffective as my Silhouette measure with the strongest performing K-means model was 0.301 and not strong enough for a significant outcome. The clusters were not even in distribution and two of the quantitative fields I was relying on proved to be the least involved in the cluster segmentation.

I decided to run a refined model using four inputs: Hours per week, Race, Sex, and Capital Gain. My cluster groupings were disproportionate, but yielded a 0.716 silhouette score with the Kohonen model yielding a statistically significant outcome.



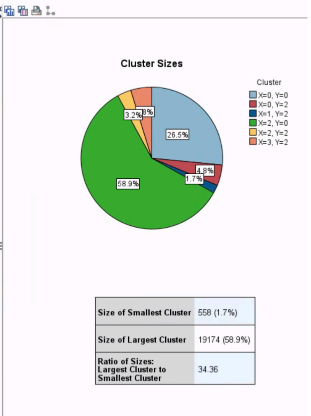
Sort by: Use ● Ascending ○ Descending View: Training set

| Use? | Graph | Model | Build Time (mins) | Silhouette | Number of Clusters | Smallest Cluster (N) | Smallest Cluster (%) | Largest Cluster (N) | Largest Cluster (%) | Smallest/ Largest | Importance |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ☑ | | Koh... | < 1 | 0.716 | 6 | 558 | 1 | 19174 | 58 | 0.029 | 1.0 |
| □ | | K-m... | < 1 | 0.696 | 5 | 574 | 1 | 19174 | 58 | 0.030 | 1.0 |
| □ | | Two... | < 1 | 0.673 | 3 | 4922 | 15 | 19020 | 58 | 0.259 | 0.952 |

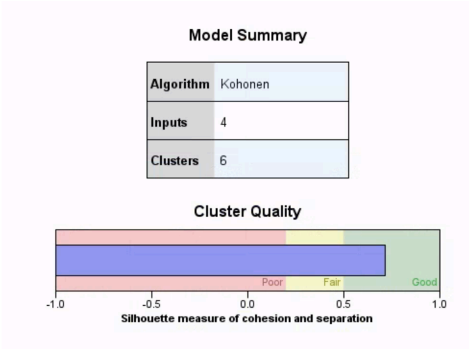


Clusters

| Cluster | X=2, Y=0 | X=0, Y=0 | X=0, Y=2 | X=3, Y=2 | X=2, Y=2 |
|---|---|---|---|---|---|
| Label | | | | | |
| Description | | | | | |
| Size | 58.9% (19174) | 26.5% (8642) | 4.8% (1571) | 4.8% (1569) | 3.2 (104 |
| Inputs | hours_per_week 42.67 | hours_per_week 36.30 | hours_per_week 36.62 | hours_per_week 40.00 | hours_per_week 41.66 |
| | race White (100.0%) | race White (100.0%) | race Black (99.0%) | race Black (100.0%) | race |
| | sex Male (100.0%) | sex Female (100.0%) | sex Female (100.0%) | sex Male (100.0%) | sex Male (100.0%) |
| | capital_gain 1,368.67 | capital_gain 573.61 | capital_gain 512.00 | capital_gain 702.45 | capital_gain 1,549.05 |



| | |
|---|---|
| Size of Smallest Cluster | 558 (1.7%) |
| Size of Largest Cluster | 19174 (58.9%) |
| Ratio of Sizes: Largest Cluster to Smallest Cluster | 34.36 |



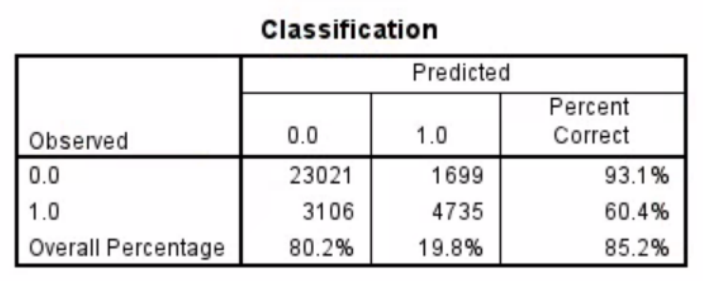
Model Summary

| | |
|---|---|
| Algorithm | Kohonen |
| Inputs | 4 |
| Clusters | 6 |

The lowest and highest number of hours worked per week were the most significant in segmentation. I found that using too many categorical variables in my original model was fragmenting my cluster attempts too much, but I was unable to create a fairly distributed set of groups while maintaining a significant silhouette score.

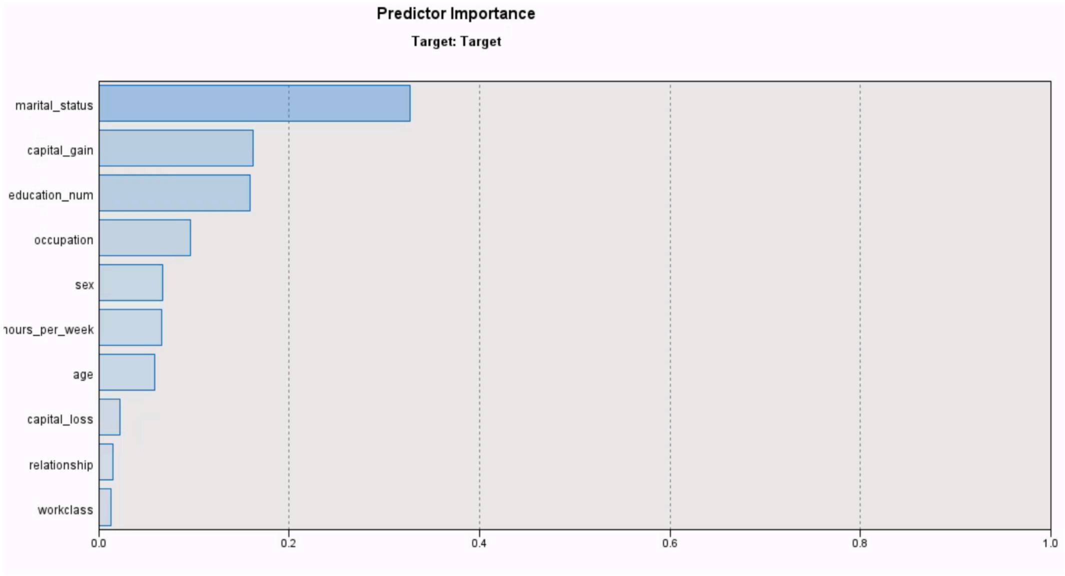
*Part III*: *Model Development*

I chose to use a logistical regression analysis due to the greater number of categorical variables in the data set and our dependent variable being binomial. This will allow us to better predict the outcome of the target variable based on the observations within the data set. I used all fields as inputs for the first run.



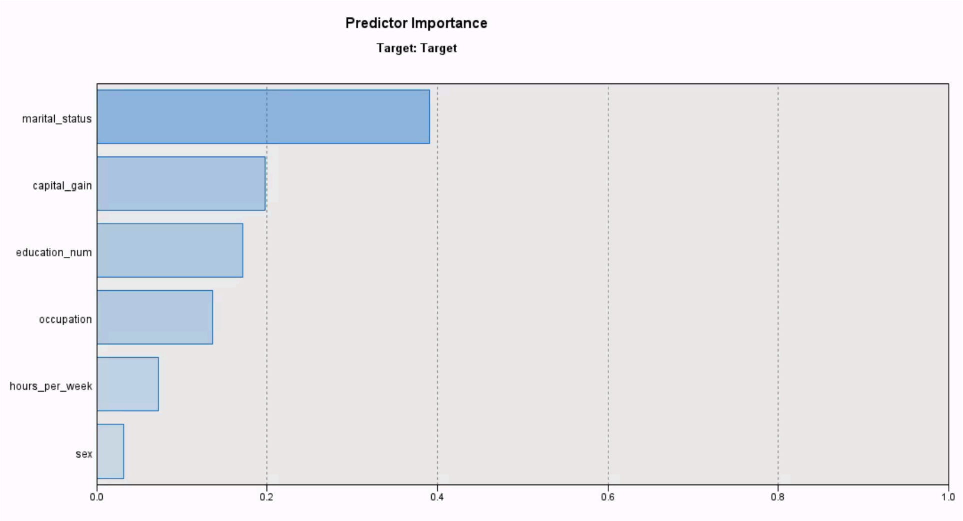
Classification

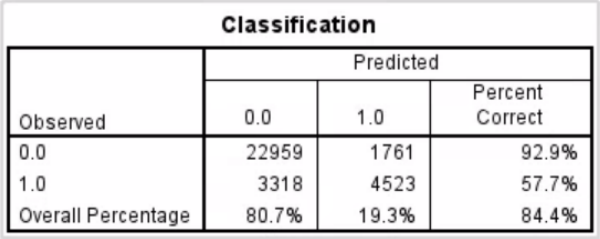
| | Predicted | | |
|---|---|---|---|
| Observed | 0.0 | 1.0 | Percent Correct |
| 0.0 | 23021 | 1699 | 93.1% |
| 1.0 | 3106 | 4735 | 60.4% |
| Overall Percentage | 80.2% | 19.8% | 85.2% |



The first run proved to be very strong with 93.1% predicted as non-targets correctly and 60.4% correctly identified as targets. However, there were multiple fields that were not noted as important predictors, and others that did not provide a significant predictor importance.

I ran a refined model with the first six variables listed in the above predictor table.





Classification

| Observed | Predicted | | |
|---|---|---|---|
| | 0.0 | 1.0 | Percent Correct |
| 0.0 | 22959 | 1761 | 92.9% |
| 1.0 | 3318 | 4523 | 57.7% |
| Overall Percentage | 80.7% | 19.3% | 84.4% |

The second run proved to have a slightly less significant efficacy and it may have been due to the mix of inputs and my omission of many categorical variables.  The percent of predictive accuracy is ok but could use work.

### *Conclusion*

Overall, the predictive accuracy of the fields within this data set is capable of indicating whether or not an individual will make >50k in a year.