Jonathan Pedoeem REU UNCW 2018

Professor Chen 05/25/2018

**Week One Report**

1. **Introduction**

Week one of the REU program focused on an introduction to the science of data mining, regression, and classification. There were three projects that were assigned using different subsets of the Morph-II database.

Morph-II is a database of police mug-shots. In addition to the actual photo, other pieces of data are given such as gender, race, DOB, and BIF (Bio-Inspired-Features).

The projects assigned include extracting relationships between race, gender, age, and BIF data, and building models to predict one of these features given another.

1. **Methodology**

The data that was provided was not fully ready for performing the necessary analysis. Age, and gender were not easily accessible as they were provided as part of the name of the image. Therefore, the first step was to take the given information and extract age and gender into their own columns in the data-frame.

In addition to basic cleaning, there was also cleaning that that needed to be done due to the general dirtiness of the data. For example, one of the fields provided in one of the datasets is the number of pictures a person has in the database. A naïve assumption would be that every user starts from picture zero. This turned out to be wrong. To get an accurate count of picture for each person, this detail needs to be taken into consideration.

Another issue with the data is that some entries do not have a consistent age, gender or race. This issue was dealt by Bingham et al., in which they released a cleaned version of Morph-II.

Following cleaning the data and performing graphical and numerical summaries, the data was used to generate both regression and classification models.

Regression is statistical method to estimate relationships between variables. This allows a user to use known data to predict the outcome of a new datapoint. There are several different methods that can be used to come up with such a model. Linear, quadratic, and cubic models were used to predict the age of an entry based on a BIF value. In the case of the linear fit, a multi-dimensional linear fit, to twenty BIF values instead of one, was also attempted. Adjusted R2 was used to measure the goodness of fit.

While Regression creates a model to predict continuous values, Classification creates a model to predict discrete values. Using the Morph-II database the classification methods: logistic regression, linear discriminant analysis, quadratic discriminant analysis, and K near nearest neighbors were used to predict the gender of an entry based on 100-175 BIF values (dependent on which classification method was used). Accuracy (correct classifications/total entries), Sensitivity (correctly classified positives/ total positives), and Specificity (correctly classified negatives/total negatives) is used to determine the goodness of the classification model.

1. **Results**

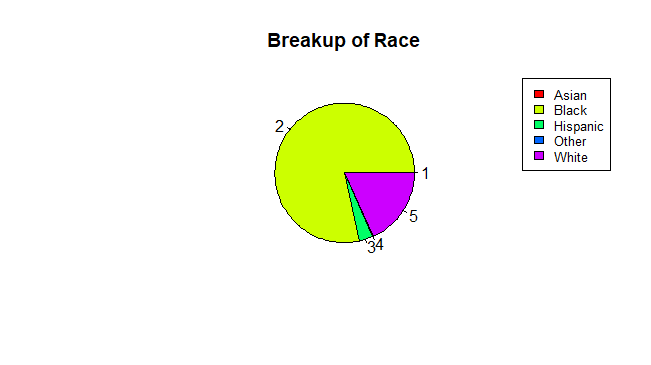
Based on the sample of 1000 entries from the Morph-II database most (84%) entries are male (see figure 2). However, the distributions of age (see figure 1) and BIF (see figure 3 and 4) values for each gender is very similar.

|  |  |
| --- | --- |
| *Figure 1: A Boxplot of Age vs. Gender for Entries (0=Female 1=Male)* | *Figure 2: A Pie Chart Representing the Breakup of Gender from the Sample Data* |
| *Figure 3: Boxplot of BIF Values for Males in the Sample Data* | *Figure 4: Boxplot of BIF Values for Females in the Sample Data* |

The distribution of BIF values is skewed to the left, a good portion of the values lie in the range 140-250 (see figure 5 and 6, table 1). The distribution of age is skewed to the right with most of the convicts are younger than 33 years of age (see figure 7 and 8, table 2).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **https://lh4.googleusercontent.com/YP9cmZjIF8w_S9ozI5pDOuAEFP0xRVpO4ctUSB3cv9uw4BklZBJG-PObLZkh33CI8gjUaZnqgrgBTEEN7iBifMjz6pwySnfoxCIi2l6mOlNjRyZS7UTRN7_0M1bRX_27ifoGfiga***Figure 5: Boxplot of all BIF Values* | |  |  | | --- | --- | | **Five Number Summary for BIF Values** | | | Min | 9.0 | | 1st. Quartile | 173.0 | | Median | 202 | | Mean | 200.1 | | 3rd Quartile | 233.0 | | Max | 255.0 |   *Table 1: A Table showing the Five Point Summary for BIF Values Gathered from the Sample Data* |
| *Figure 6: A Histogram Showing the Frequency of BIF Values for all 1000 Samples* | *Figure 7: A Histogram Representing the Frequencies of Ages in the Provided Data* |
| *Figure 8:**A Boxplot showing the Five Points for age* | |  |  | | --- | --- | | **Five Number Summary for Age** | | | Min | 16.0 | | 1st. Quartile | 23.0 | | Median | 32.0 | | Mean | 32.4 | | 3rd Quartile | 40.3 | | Max | 70.0 |   *Table 2: A Table showing the Five Point Summary for Ages*  *Gathered from the Sample Data* |

Analysis of race shows that most of arrests are Black (see figure 9). A comparison of table 3 and table 4 shows the discrepancies between the dirty Morph-II data and the clean one. The dirty one had people has several entries where in different places the same person has a different race. For table 3 the first instance of race was chosen, while for table 4 a more detailed method is used to pick the more suitable race.



*Figure 9: Breakup of Race based on Cleaned Morph-II Data*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Race by Gender on Dirty Data** | | | | | | | |  | Black | White | Asian | Hispanic | Other | Total | | Male | 8838 | 2070 | 49 | 517 | 15 | 11489 | | Female | 1492 | 632 | 3 | 27 | 5 | 2158 | | Total | 10323 | 2692 | 50 | 534 | 18 | 13617 |   *Table 3: Break up of Race and Gender on the Dirty Morph-II Data* |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Race by Gender on Cleaned Data** | | | | | | | |  | Black | White | Asian | Hispanic | Other | Total | | Male | 8829 | 2056 | 47 | 507 | 19 | 11458 | | Female | 1494 | 634 | 6 | 30 | 5 | 2169 | | Total | 10332 | 2704 | 55 | 547 | 20 | 13658 |   *Table 4: Break up of Race and Gender on the Clean Morph-II Data* |

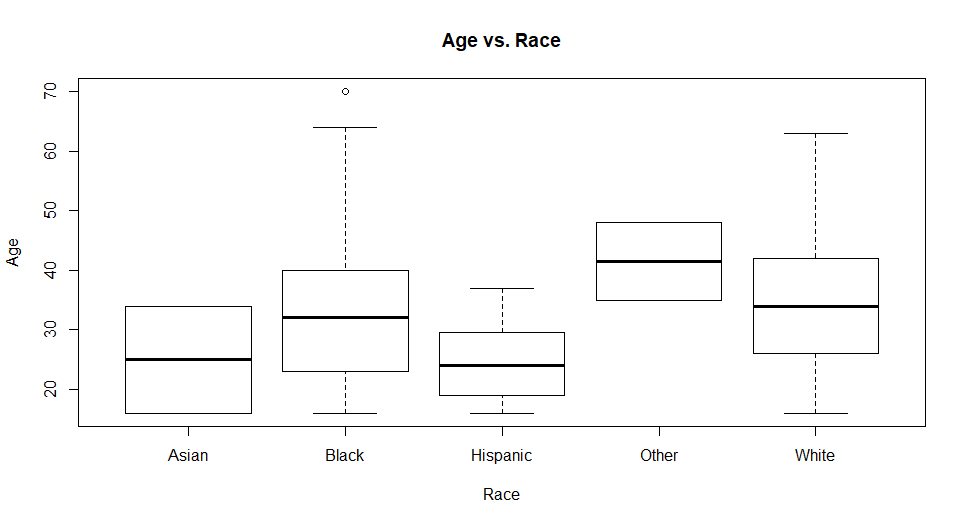
Table 5 shows the breakup of arrests with the count starting at zero. Based on this table most males and females have been arrested a total of 3 times.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Total Arrests by Gender** | | | | | | | |  | 1 | 2 | 3 | 4 | 5+ | Total | | Male | 2350 | 3606 | 1975 | 1135 | 2020 | 11086 | | Female | 478 | 712 | 352 | 172 | 360 | 2074 | | Total | 2828 | 4318 | 2327 | 1307 | 2390 | 13160 |   *Table 5: Break up Total Arrests by Gender Based on the Clean Morph-II Data* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age Groupings by Gender** | | | | | | |
|  | <20 | 20-29 | 30-39 | 40-48 | 49+ | Total |
| Male | 1968 | 3387 | 3047 | 2165 | 891 | 11458 |
| Female | 294 | 607 | 686 | 451 | 121 | 2159 |
| Total | 2262 | 3994 | 3733 | 2616 | 1012 | 13617 |

Table 6 shows the age of each of the convicts at the time of their first arrest. The data shows that most male convicts are arrested when they are in between 20-29 while most females are between 30-39. This table differs from figure 8 and table 2 as this is the age when they were first arrested and not the general distribution of ages across all arrests.

*Table 6: Break up of Age at First Arrest by Gender Based on the Clean Morph-II Data*

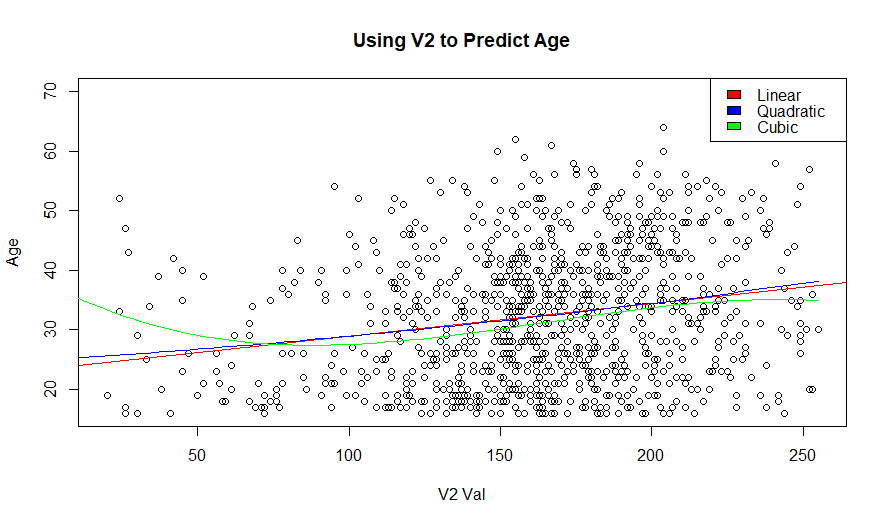
Figure 10 shows the breakup of age vs race. While there are some discrepancies between the ages of difference races, it is important to mention that Asian, Hispanic, and Other category only had a few entries, this could have skewed their distributions. A more accurate comparison would require a bigger dataset.

*Figure 10: Breakup of Age vs. Race based on Cleaned Morph-II Data*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Comparison of Regression Methods** | | | | |
|  | Multiple Linear (20 BIF Values) | Multiple Linear (810 BIF Values) | Simple Linear (V2) | Simple Quadratic  (V2) | Simple Cubic (V2) |
| Adjusted R­­2 | 0.084 | 0.6808 | 0.0446 | 0.044 | 0.02939 |

Several regression techniques were used to create a model that predicts the age of an entry based on her BIF values. Table 7 shows a comparison of the adjusted R2  values of each. Linear Regression using 810 BIF values performed the best, achieving an adjusted R2 of 0.6808. Figure 11 shows the simple linear, quadratic and cubic fits on the actual data.

*Table 7: Comparison of Regression methods using BIF values to predict Age*

******

*Figure 11: Depiction of Simple Linear, Quadratic, and Cubic Fits*

Several different classification methods were used to predict gender based on BIF values. It is important to note that guessing all entries to be male would give a accuracy of 85.5%. None of the classification methods attempted beat that accuracy.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Comparison of Classification Methods** | | | | | | | |
|  | Accuracy  (Full) | Sensitivity (Full) | Specificity (Full) | Accuracy (80-20) | Sensitivity (80-20) | Specificity (80-20) | # of BIF Values |
| Logistic Regression | 92.4% | 95.7% | 74.5% | 81% | 83.0% | 69.0% | 175 |
| LDA | 92.9% | 97.4% | 68.8% | 82.5% | 84.8% | 31.0% | 175 |
| QDA | 100.0% | 100.0% | 100% | 85.5% | 100.0% | 0% | 150 (Full)/100 (80-20) |
| KNN (K=3) | – | – | – | 84.5% | 93.0% | 66.0% | 100 |

*Table 8: Comparison of Classification methods using BIF values to predict Gender*

1. **Conclusion**

The preliminary work done on the database offers a promising start to extracting relevant data from the Morph-II database.While some work, like classification has some ways to go the database promises to be useful in providing a good training ground to make useful inferences on age, gender, and race.