Case Type vs Race

Jaden Wilkins, jwilkins@bellarmine.edu

ABSTRACT
Each year, hundreds of thousands of children go missing in the United States. It's no secret that age, sex, and race can play a role in a child's vulnerability, so I wanted to run a program that would predict what type of case we would have based on those three variables. The dataset comes with close to 20 more columns and over 2,000 entries, so I will have lots of data to choose from.

I. INTRODUCTION

The output variable in my model is "casetype". My goal is to create a predictive model that can predict how kids go missing based on their age, sex, and race. I used logistic regression to make my model. There are 5 classifications of cases: Endangered Runaway, Family Abduction, Lost Injured Missing, Non Family Abduction, and Section 5779. I changed the categorical casetype variable into casetype_num where I set: {: Endangered Runaway: 1, Family Abduction: 2, Lost Injured Missing: 3, Non Family Abduction: 4, and Section 5779: 5}

II. BACKGROUND

The dataset was found on data.world. The dataset contains cases from mostly the United States, but a few in Mexico and Canada. The data is publicly available information from the National Center for Missing and Exploited Children. (NCMEC)

III. EXPLORATORY ANALYSIS

In the beginning of my project, there were 19 columns with 2834 entries or rows. The dataset came in only object and int64 variable types.

In this summary, provide the data types of your columns (in a table) and then rather than providing tabular statistics and plots for each variable, provide only statistics and plots that seem unusual. For example, if one or two variables have significant missing values or the distribution of the variable is skewed or looks unusual note that. Provide the unusual statistics or plots in this section. Provide any other appropriate plots (e.g. correlation matrix, heatmaps, bar charts, etc.) that you deem necessary.

Table 1: Data Types

Variable Name	*Data Type*
missingreporteddate	Object --→ datetime
birthdate	Object -→ datetime
Sex object	Object ---> int32
Age (not included)	Float
missingfromdate	Object --→ datetime

Figure 1) Race vs. Age

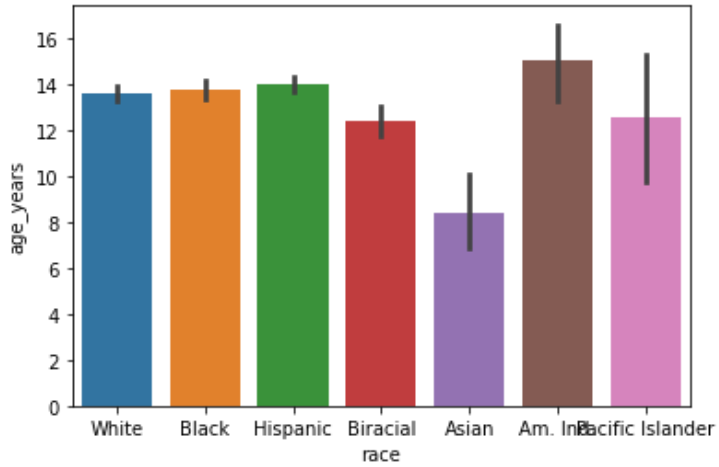


Figure 2) Race vs. Casetype (numerical)

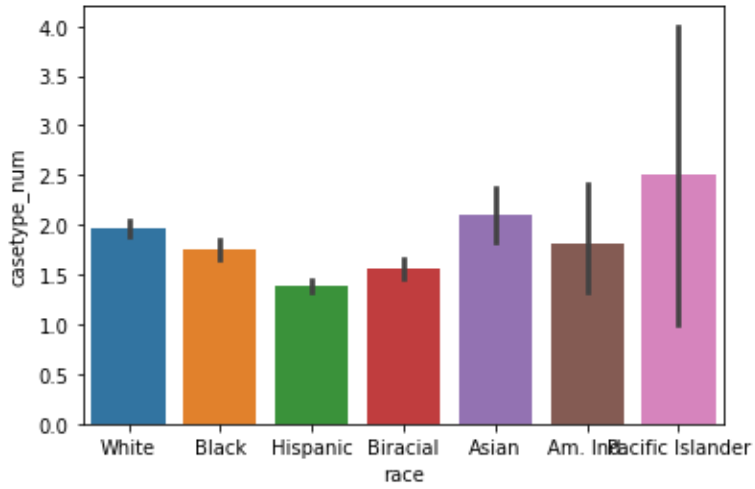


Figure 3) Correlation Heatmap

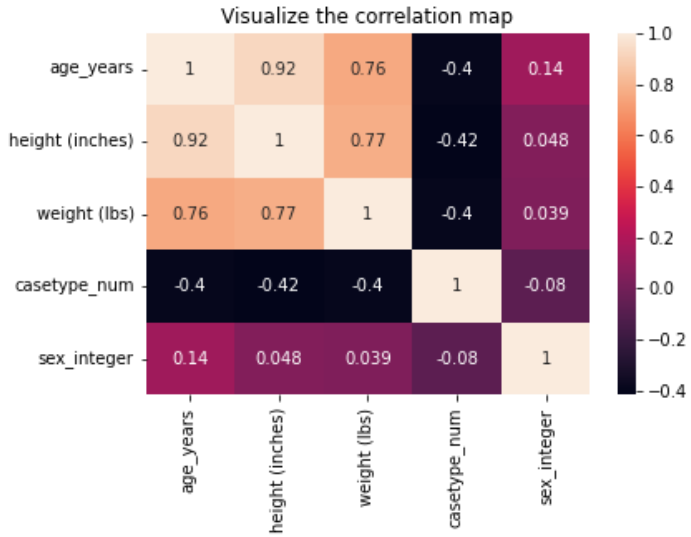


Figure 4) Pie Chart of Case Types

To visualize the percentage of each case type.

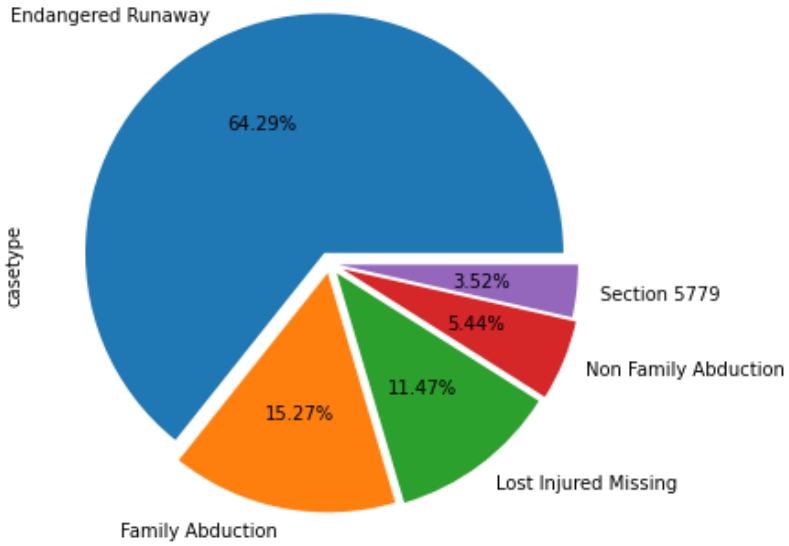
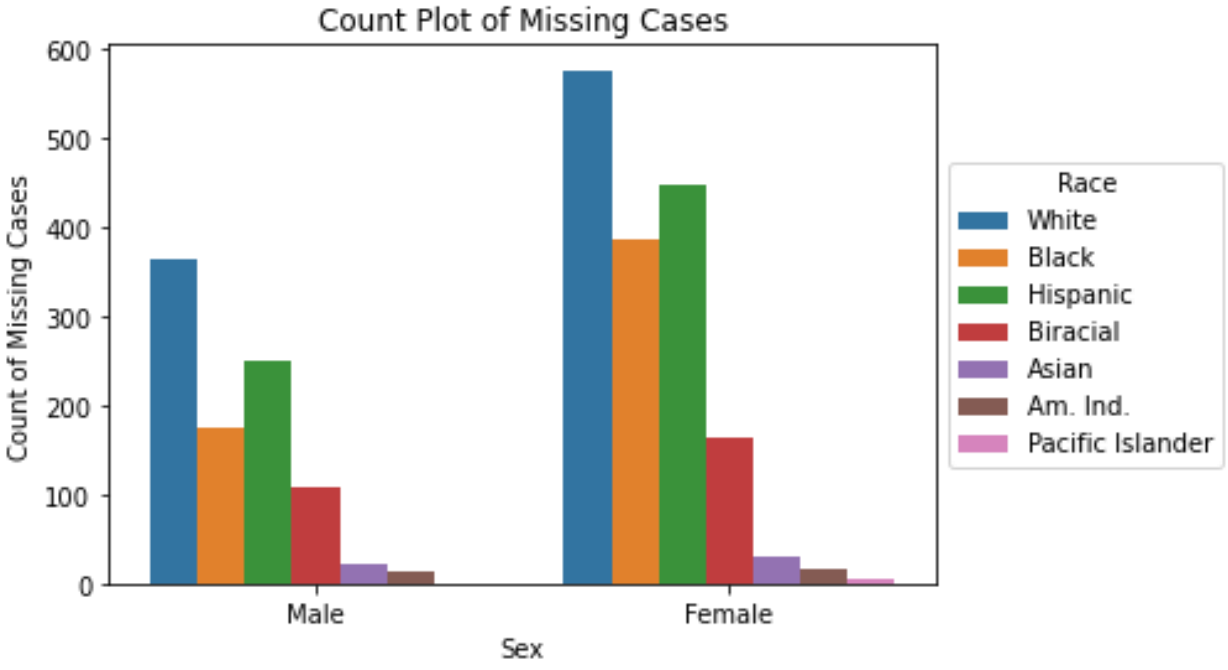


Figure 5) Count Plot of missing Cases



IV. METHODS

In this section, describe how you prepared the data for your model and performed multiple experiments using different parameters for the model(s).

A. Data Preparation

The dataset only came with null values in a few columns. I omitted the rows with the missing values because they were categorical and I was not losing a lot of data relative to the size of the dataset. Interestingly, there was no “age” column in the dataset so I converted the “birthdate” and “missingfromdate” to datetime cells and subtracted the two to get “age”. I had to change “race” into a number so I numbered them in order by population starting with “White” at 1. I also had to change sex to a binary variable where male was 0 and female was 1.

B. Experimental Design

In [38]:
```
y_pred=logreg.predict(X_test)
print (X_test) #test dataset
print (y_pred) #predicted values
```

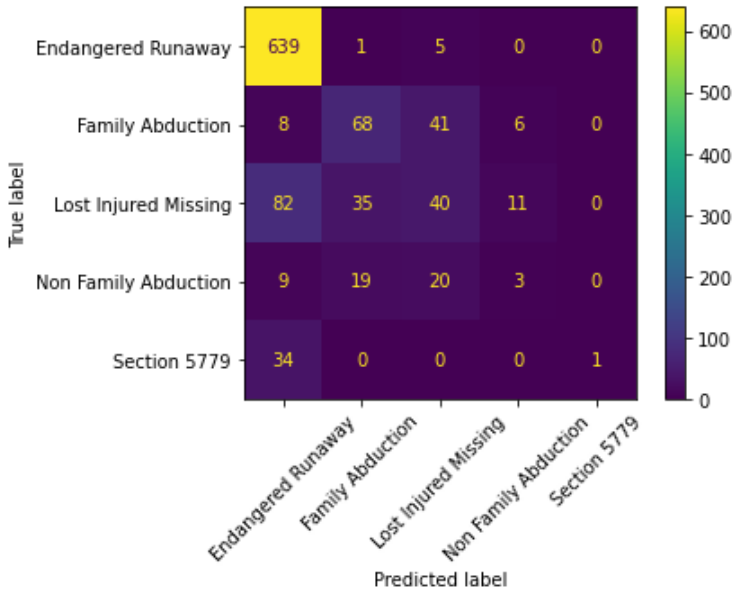
```
      race_num  sex_integer  age_years
1185         1            1        6.9
1459         1            1       16.3
1706         1            0       16.5
1711         2            1       17.0
1051         5            0        2.7
...        ...          ...        ...
1757         2            1       14.9
2107         3            0       17.8
563          2            0        2.0
1408         2            1       14.0
239          3            1        8.7

[1022 rows x 3 columns]
['Lost Injured Missing' 'Endangered Runaway' 'Endangered Runaway' ...
 'Family Abduction' 'Endangered Runaway' 'Lost Injured Missing']
```

In [43]:
```
from sklearn import metrics
from sklearn.metrics import classification_report
print('Accuracy: ',metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy:  0.7348336594911937
```

```
In [55]: disp=ConfusionMatrixDisplay(confusion_matrix=cm,
                                     display_labels=logreg.classes_)
         disp.plot()
         plt.xticks(rotation=45)
         plt.show()
```



C. Tools Used

Describe all of the software tools you used to perform your data preparation and model implementation. For example:

The following tools were used for this analysis: I used Python via Jupyter Notebook on my Windows HP laptop. In addition to Python I used numpy, pandas, matplotlib, and seaborn.

V. RESULTS

A. Classification Measures/ Accuracy measure

In [38]:
```
y_pred=logreg.predict(X_test)
print (X_test) #test dataset
print (y_pred) #predicted values
```

```
      race_num  sex_integer  age_years
1185         1            1        6.9
1459         1            1       16.3
1706         1            0       16.5
1711         2            1       17.0
1051         5            0        2.7
...        ...          ...        ...
1757         2            1       14.9
2107         3            0       17.8
563          2            0        2.0
1408         2            1       14.0
239          3            1        8.7

[1022 rows x 3 columns]
['Lost Injured Missing' 'Endangered Runaway' 'Endangered Runaway' ...
 'Family Abduction' 'Endangered Runaway' 'Lost Injured Missing']
```

B. Problems Encountered

I had a lot of trouble with the regression because I had a mix of binary and non-binary variables. Eventually I was able to find a way to make it work although it did not go as I originally planned. There were supposed to be another 2 X columns, height and weight, that would play a role.

Additionally, the first dataset I picked ended up not being as compatible as I thought, so I had to switch.

C. Limitations of Implementation

One limitation my dataset had was presenting race and sex as numbers and not categorical variables. While it was necessary for the model, it admittedly made things complicated when I tried to run the prediction.

D. Improvements/Future Work

In the future, learning how to deal with non-binary variables would help me the most in terms of what to improve on. I think I had a strong idea for something to investigate but I couldn't quite piece it together the way I imagined.

VI. CONCLUSION

Overall, I think I had a strong idea for a regression model. With decent accuracy at 73% I am proud of the outcome but I could have done more on the regression.

https://data.world/jamesgray/missing-children-in-the-us