Problem Statement

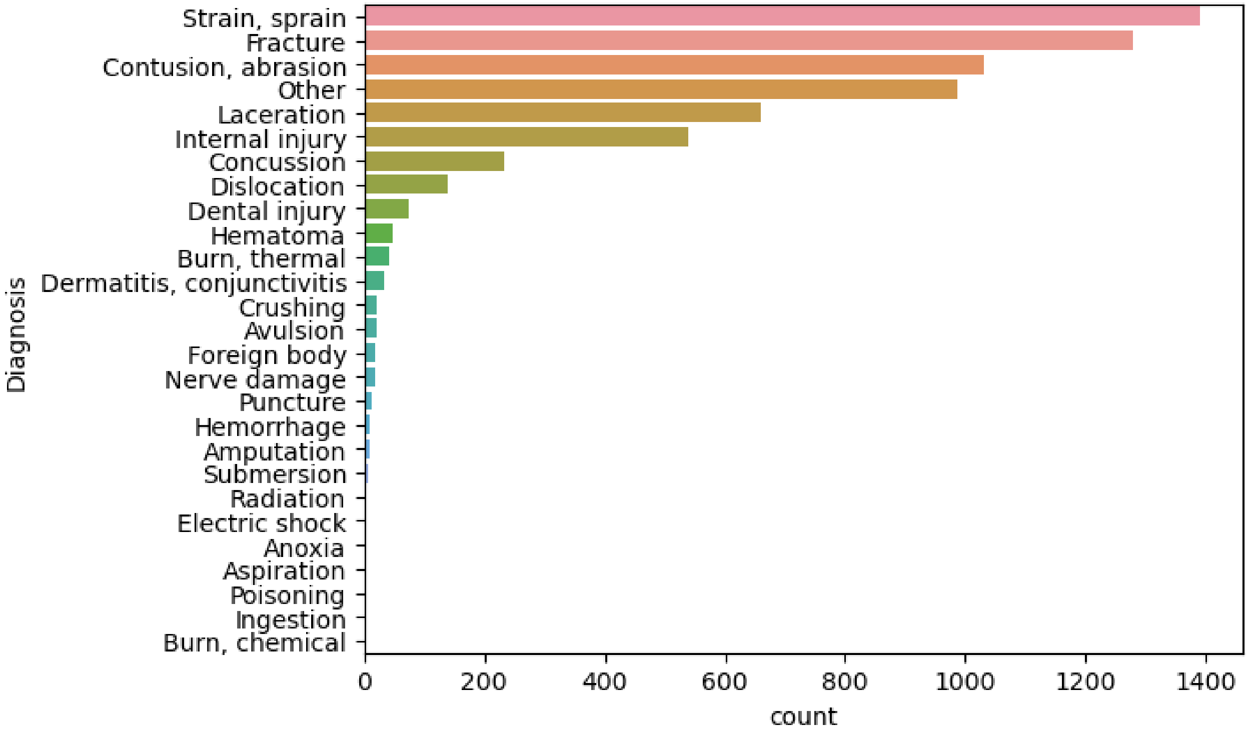
How can we predict a common injury diagnosis to allow the park to safeguard against these injuries within the next year by allocating medical personnel, procedures, supplies and warning signs properly and increase patron safety?

Data Wrangling

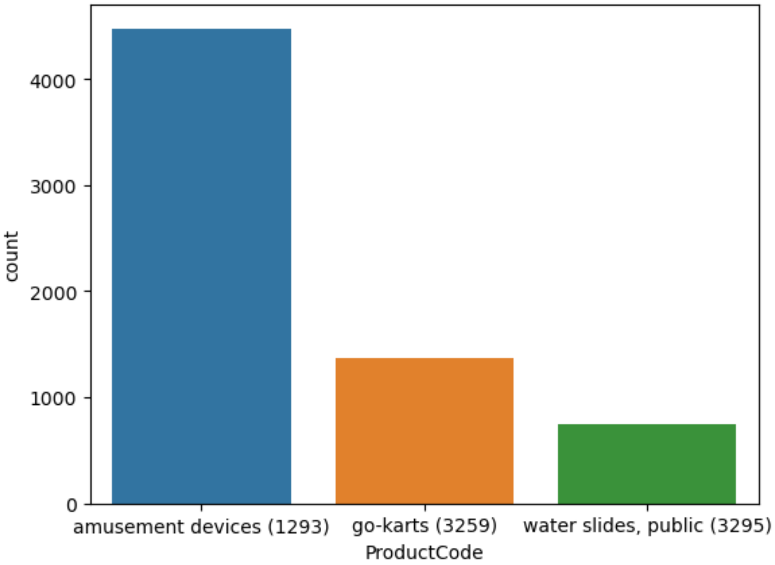
Sourcing our data from the 2018 NEISS Analysis, we have data from hospital records from 2013-2017 for amusement ride related injuries. We can use this data to analyze patterns and predict amusement ride and device related injuries. For our purposes of this report, we will be using this data and predict whether an injury is a common injury or a non common injury. With this information, amusement parks can determine some guidelines and safety procedures to treat these injuries quickly or prevent them from happening entirely through ride procedures.

Exploratory Data Analysis

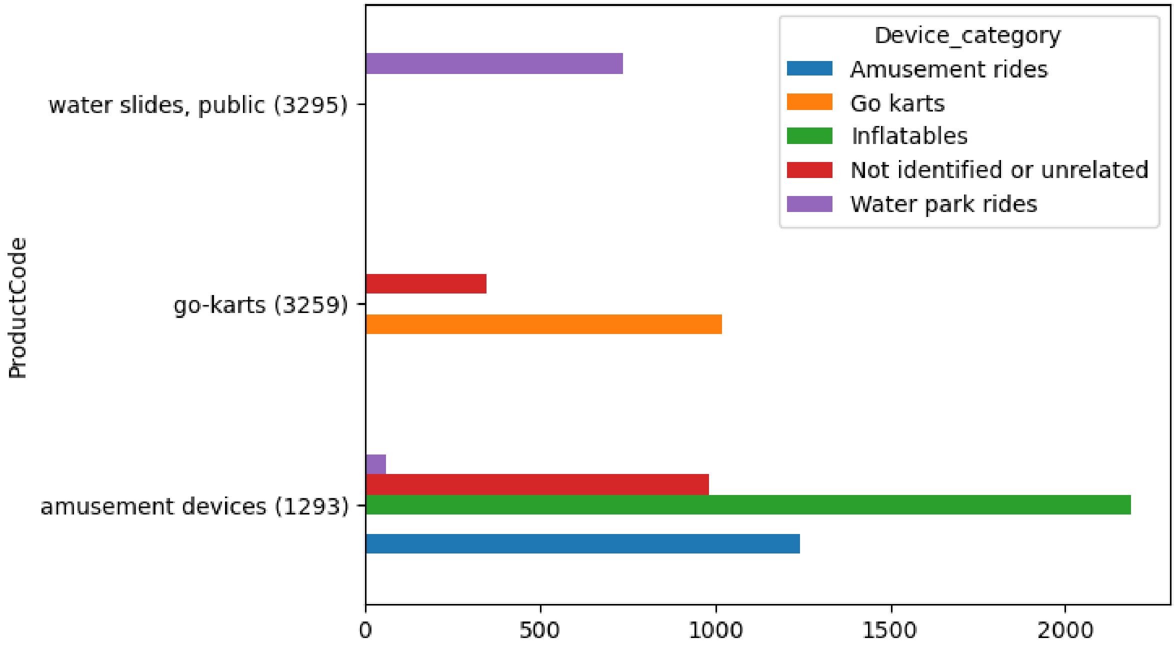
After reviewing the data provided and cleaning the data to make sure it is valid for our models, we found that there are three main common diagnoses: Strains/Sprains, Fractures, and Contusions.



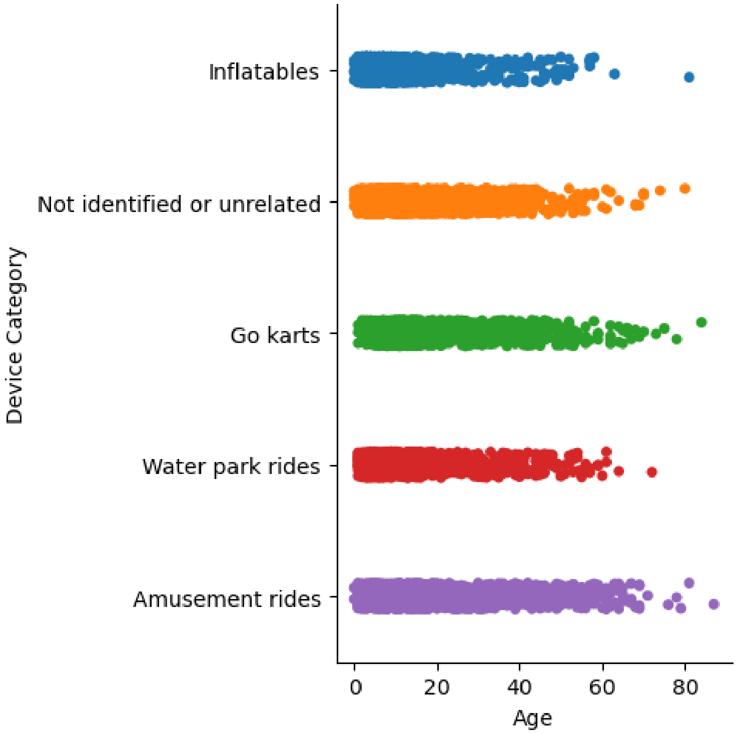
Looking at the types of products where injuries occurred most often, we see that Amusement Devices appear to cause the most injuries:



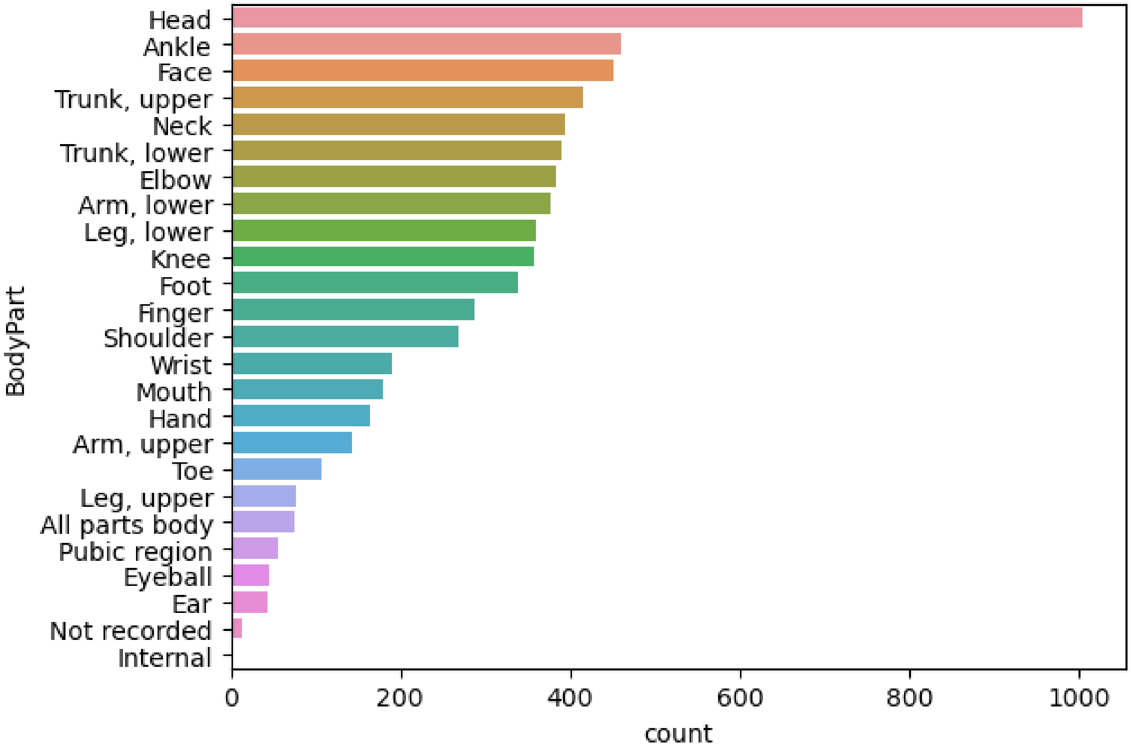
Of the device categories that make up the Amusement devices, we can see that it includes Inflatables which is the largest of any of the devices.



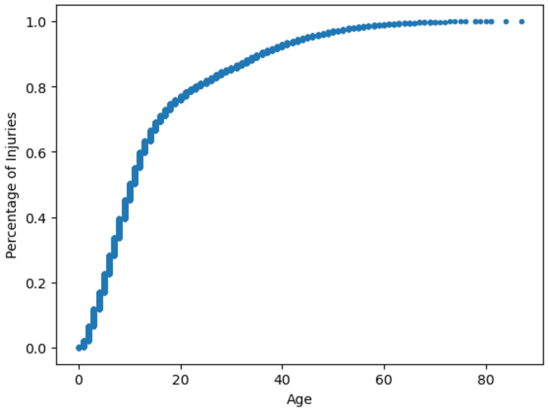
We also see that for people injured on Inflatables, the population tends to skew younger, especially compared to the other device categories.

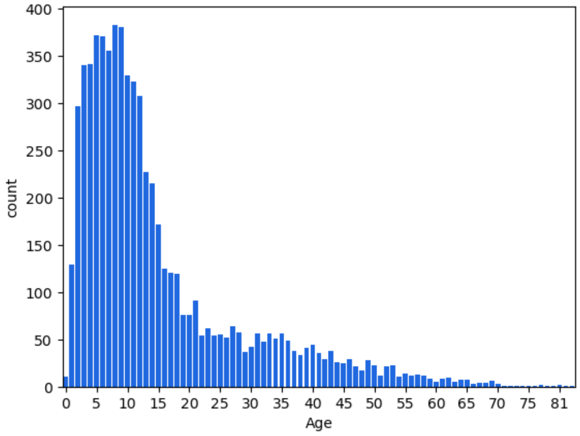


For the most commonly injured Body Part, we see that the Head is more than double the next most common injury area.

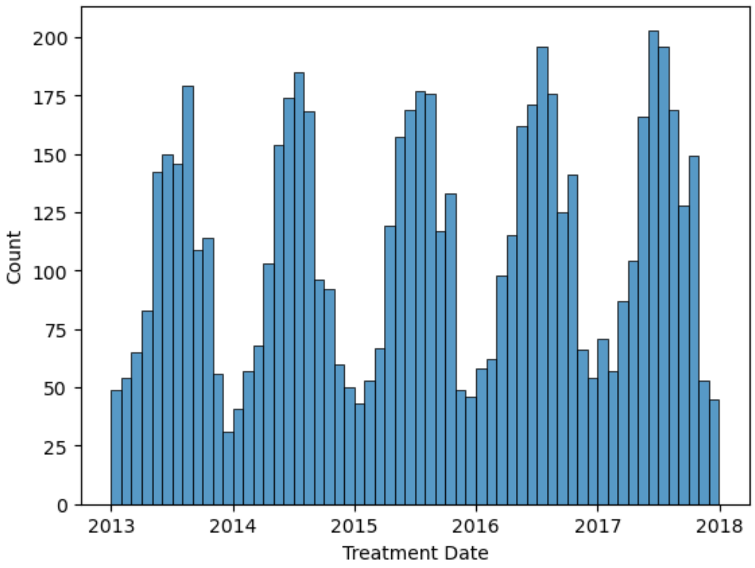


We also see that around 80% of the injuries are people under the age of 20:





We see some seasonality in the treatment date as well where there are a higher number of injuries in the summer months.

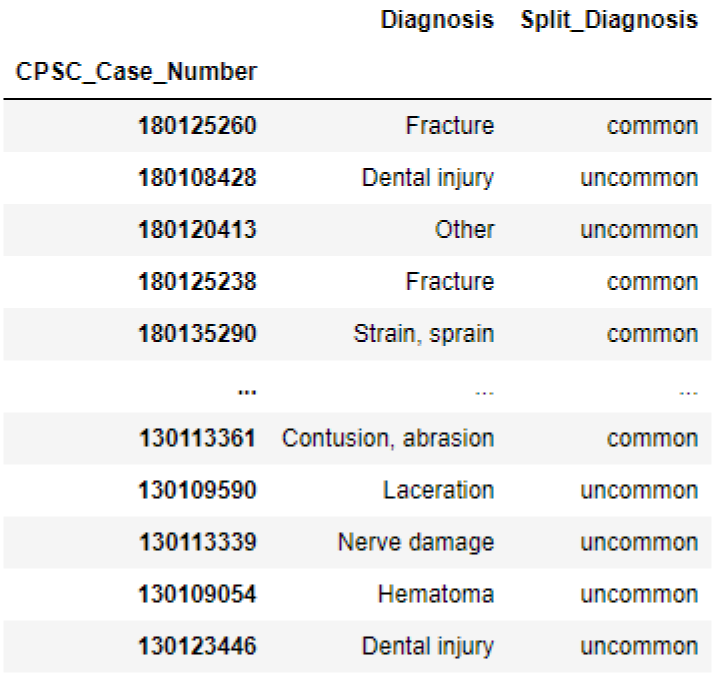


Making note of these observations, we can now move into creating our models.

Model Preprocessing with feature engineering

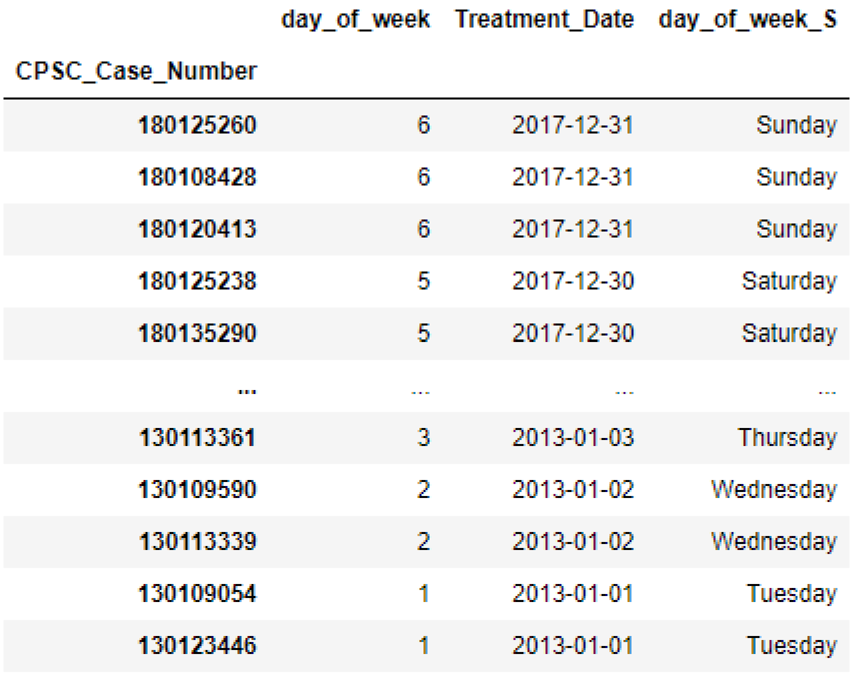
Our first step was to encode dummy variable columns so that all the categorical columns are split out as 0 or 1 as opposed to strings. We dropped the first column as well so there was no duplicate information affecting our data. Then we converted our Diagnosis column that had all the diagnosis of all the injuries and created the Split Diagnosis column which separated the Diagnosis into two categories: Common and Uncommon. The Common injuries were the ones that we found during our Exploratory Data Analysis which includes “Strains/Sprains”, “Fractures”, and “Contusions”.

Everything else is considered an Uncommon Diagnosis. We then created dummy variable columns, as well as made the Common Split Diagnosis our target column (1 for Common, 0 for Uncommon).



	Diagnosis	Split_Diagnosis
CPSC_Case_Number		
180125260	Fracture	common
180108428	Dental injury	uncommon
180120413	Other	uncommon
180125238	Fracture	common
180135290	Strain, sprain	common
...	...	...
130113361	Contusion, abrasion	common
130109590	Laceration	uncommon
130113339	Nerve damage	uncommon
130109054	Hematoma	uncommon
130123446	Dental injury	uncommon

We also created some extra features for Day of the Week (Monday-Sunday signified as 0-6) so that can be included as a feature in our model.



	day_of_week	Treatment_Date	day_of_week_S
CPSC_Case_Number			
180125260	6	2017-12-31	Sunday
180108428	6	2017-12-31	Sunday
180120413	6	2017-12-31	Sunday
180125238	5	2017-12-30	Saturday
180135290	5	2017-12-30	Saturday
...	...	...	...
130113361	3	2013-01-03	Thursday
130109590	2	2013-01-02	Wednesday
130113339	2	2013-01-02	Wednesday
130109054	1	2013-01-01	Tuesday
130123446	1	2013-01-01	Tuesday

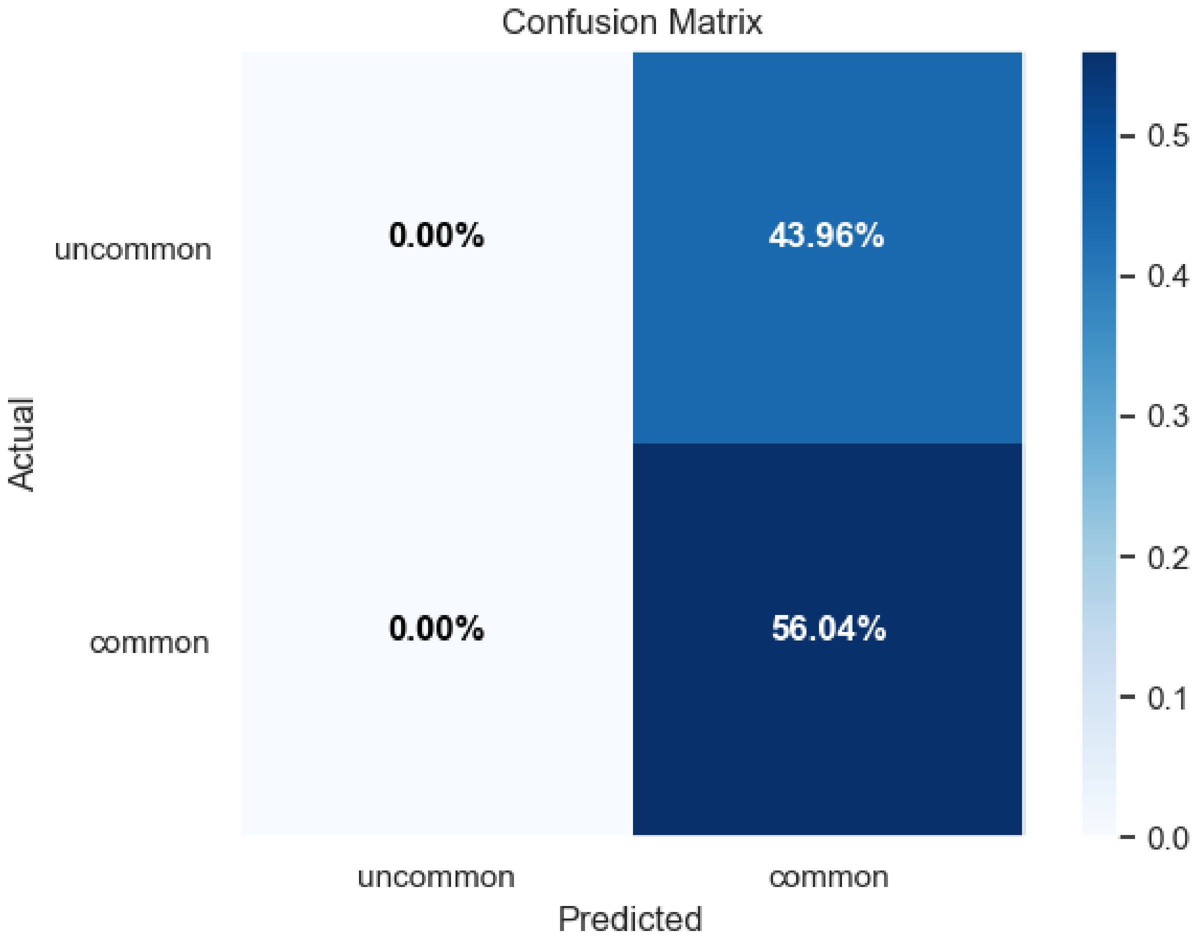
Once these features were created, we split the data into train and test sets and then scaled all the numerical values. Now we are ready for our models.

Modeling

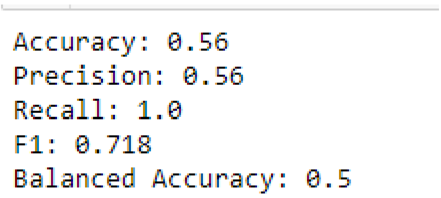
The main two ways we are checking our results of the models are through confusion matrices and scoring statistics of Accuracy, Precision, Recall, F1, and Balanced Accuracy.

Baseline Model

For our first model, we took the mean value of the training set for Common Diagnosis, and then used that as our initial model to assume everything was a Common Diagnosis (56.26% in the train set). Doing this resulted in correctly predicting 56.04% of the test set.



We see similar Precision and Accuracy scores, but a perfect recall due to having no False Negatives.

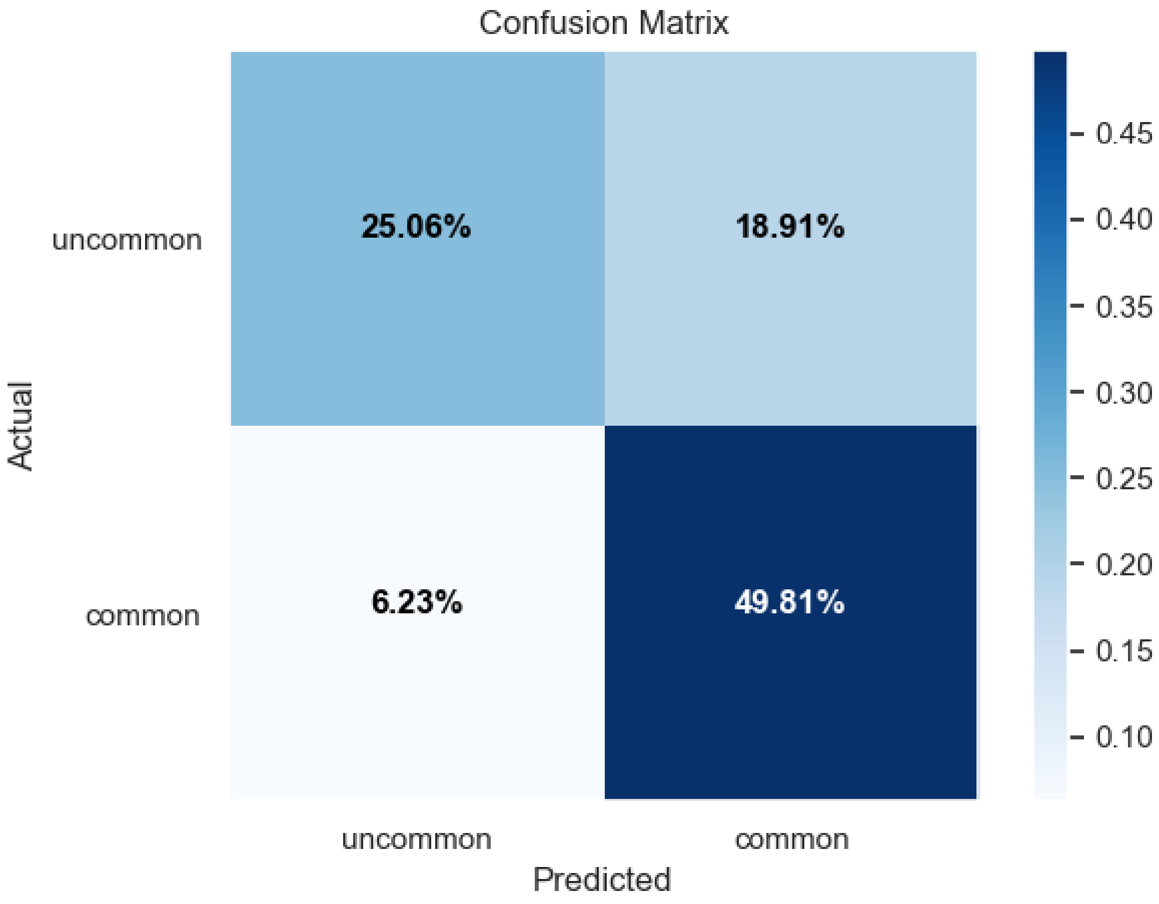


```
Accuracy: 0.56
Precision: 0.56
Recall: 1.0
F1: 0.718
Balanced Accuracy: 0.5
```

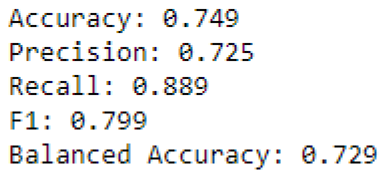
This will be the baseline that we can compare to all other models and find improvements between the models.

Logistic Regression

Next we did a Logistic Regression model and we see we were slightly worse at predicting the Common Diagnosis (still a majority of the actual common diagnosis), but much better at predicting the uncommon diagnosis.



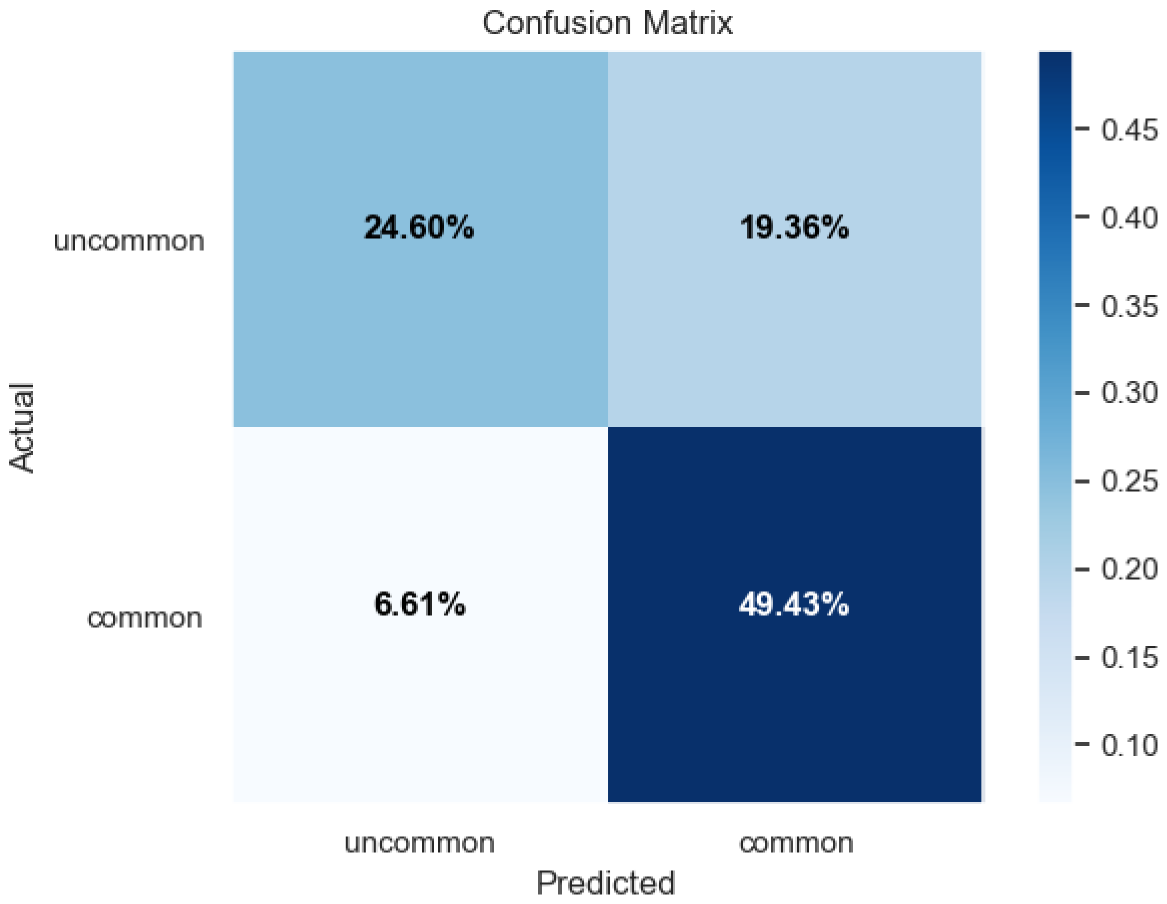
As for the scores, we see that vast improvement as well with increases in everything except Recall which was perfect in the Baseline Model.



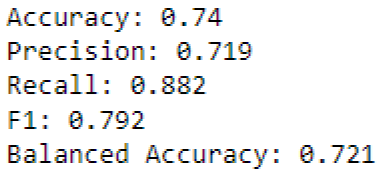
```
Accuracy: 0.749
Precision: 0.725
Recall: 0.889
F1: 0.799
Balanced Accuracy: 0.729
```

Random Forest

Lastly we created a Random Forest model. Compared to the Logistic Regression model, it is slightly better at predicting uncommon diagnosis, and slightly worse at predicting common diagnosis, but they are so close together, that they get very similar results.



We can see that with the scores where all the metrics are slightly worse than the Logistic Regression model.



```
Accuracy: 0.74
Precision: 0.719
Recall: 0.882
F1: 0.792
Balanced Accuracy: 0.721
```

Recommendation and Conclusion

After reviewing the three models, we found that the best model is the Logistic Regression Model based on our testing. It has the best overall scores and is the best model to predict the Common Diagnosis. Using this model, we can predict the types of injuries to occur given different Ages, Genders, Ride type and categories, locations, day, month and year. We can then use this model on each ride to determine appropriate safety precautions and procedures for the park to implement.

Future Scope of Work

The current hyperparameters have not been tuned so further tuning can be done to improve model results further. Also, it may be helpful to further breakdown the common injuries and non common injuries to get a better idea of the specific types of injuries that could occur on particular ride types. This can allow us to gain more insights into the injuries that can happen and be better prepared in case something happens.