Hero or Villain?
Predicting Superhero Alignment

By Jason Masurovsky
GWU Data Analytics and Visualization Bootcamp
January 2021

Topic Choice

- Comics have been around since 1930s
- In the past decade, DC and Marvel have skyrocketed film adaptations
- Recent years, datasets on comic books and characters have been uploaded and open to the public
- Few data viz projects have been made and not many have focused on using machine learning

Research Questions

- Can we classify a superhero as good or evil based off their physical characteristics, superpowers, and abilities?
- Is there a bias in the dataset?
- Which features are most influential?

Dataset

- All the datasets were found on Kaggle
- Kaggle is webportal for the data science community
- Hosts a wide range of datasets and tooltips as open data
- Source data: https://www.kaggle.com/dannielr/marvel-superheroes
- “Characters_stats.csv”, “marvel_characters_info.csv”, and “superheroes_power_matrix.csv”

Data Cleaning

- Missing values
  - Alignment
- Duplicates
  - Names
- Convert height, weight to feet and lbs.
  - Create a BMI field
- Delete any unnecessary fields
- 517 characters in final dataset

In [2]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
#Read in datasets
ability_stats = pd.read_csv("charcters_stats.csv")
ability_matrix = pd.read_csv("superheroes_power_matrix.csv")
info_df = pd.read_csv("marvel_characters_info.csv")
```

In [4]:
```
ability_stats.head(10)
```

Out[4]:

	Name	Alignment	Intelligence	Strength	Speed	Durability	Power	Combat	Total
0	3-D Man	good	50	31	43	32	25	52	233
1	A-Bomb	good	38	100	17	80	17	64	316
2	Abe Sapien	good	88	14	35	42	35	85	299
3	Abin Sur	good	50	90	53	64	84	65	406
4	Abomination	bad	63	80	53	90	55	95	436
5	Abraxas	bad	88	100	83	99	100	56	526
6	Adam Monroe	good	63	10	12	100	71	64	320
7	Adam Strange	good	1	1	1	1	0	1	5
8	Agent 13	good	1	1	1	1	0	1	5
9	Agent Bob	good	10	8	13	5	5	20	61

In [5]:
```
ability_matrix.head(10)
```

Out[5]:

	Name	Agility	Accelerated Healing	Lantern Power Ring	Dimensional Awareness	Cold Resistance	Durability	Stealth	Energy Absorption	Flight	...	Web Creation	Reality Warping	Odin Force	Symbiote Costume	Speed Force
0	3-D Man	True	False	False	False	False	False	False	False	False	...	False	False	False	False	False
1	A-Bomb	False	True	False	False	False	True	False	False	False	...	False	False	False	False	False

In [11]:
```
#Check for NAs
df_merged.isna().sum()
```

Out[11]:
```
Name            0
Alignment       3
Intelligence    0
Strength        0
Speed           0
Durability      0
Power           0
Combat          0
Total           0
ID              0
Gender          0
EyeColor        0
Race            0
HairColor       0
Publisher       7
SkinColor       0
Height          0
Weight          0
BMI             0
dtype: int64
```

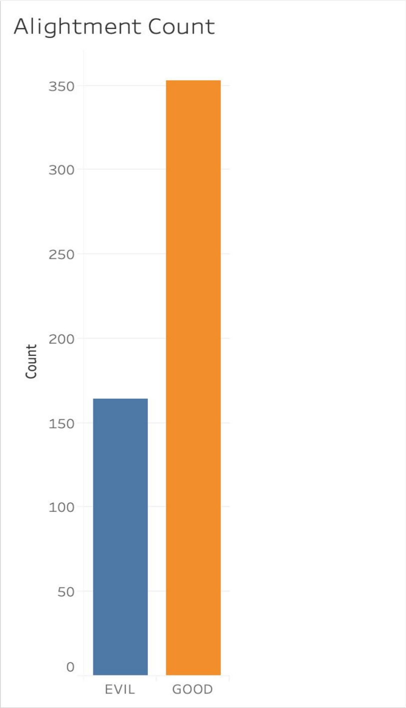
In [12]:
```
#Looked at the 3 missing values for Alignment specifically and checked online their status to fill in. All were evil.
df_merged['Alignment'] = df_merged['Alignment'].replace(np.nan, 'bad', regex=True)

#Check the neutral character to determine if can be labeled as good or bad
df_merged['Alignment'] = df_merged['Alignment'].replace('neutral', 'bad', regex=True)
df_merged.isna().sum()
```

In [13]:
```
#Convert Alignment to 1: good; 0: bad.
df_merged['Alignment'] = df_merged['Alignment'].map({'good': 1, 'bad': 0})
```

Data Analysis

- Completed in jupyter notebooks using Python and in Tableau
- Target variable: Alignment (good or evil)
- Distribution count of alignment
    - Grouped by gender, race, superpower
- Average ability stats and gen label
- Pearson Correlation on character stats such as strength, power, durability, combat, speed, intelligence
    - Determine those that are highly correlated to be removed

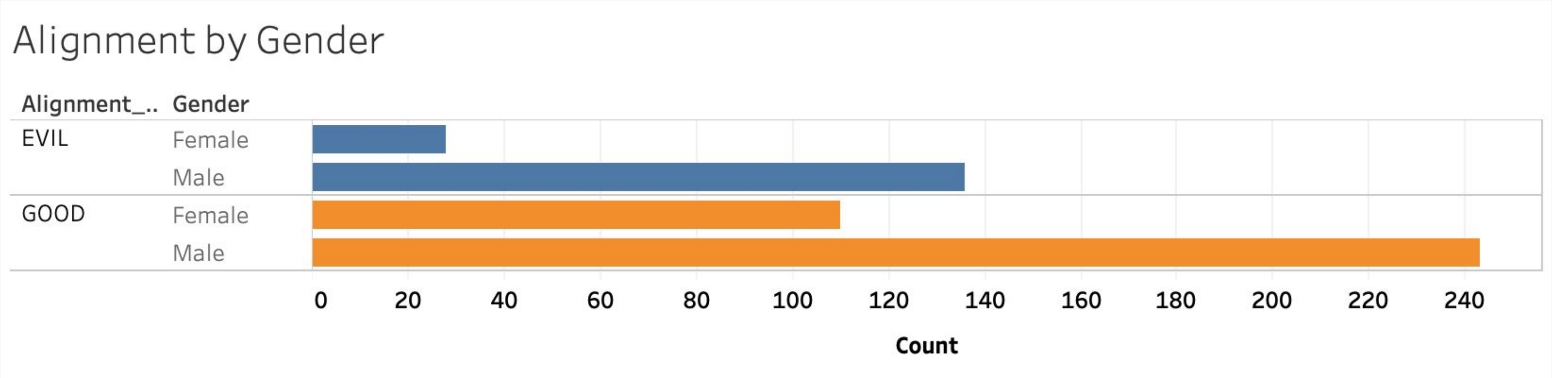
Data Analysis



Breakdown of character alignment

- 353 good characters
- 164 evil characters

Data Analysis

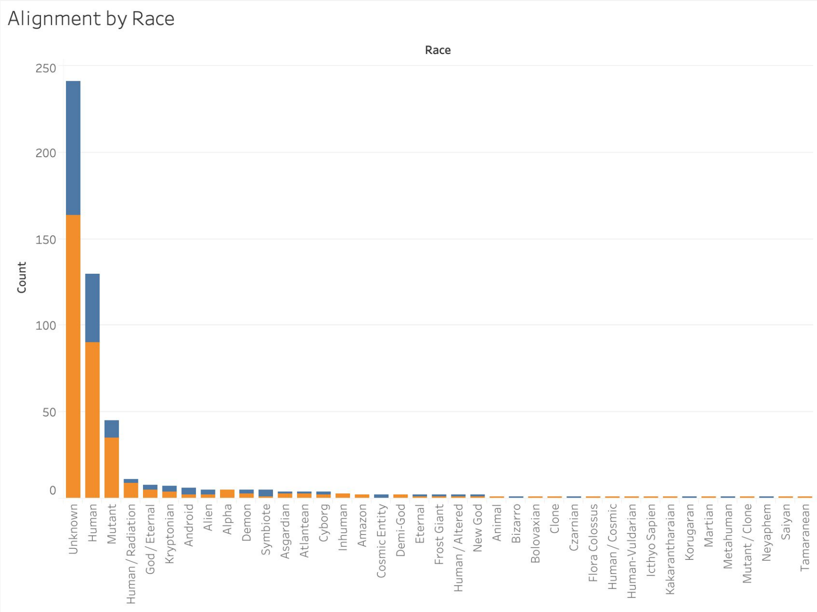


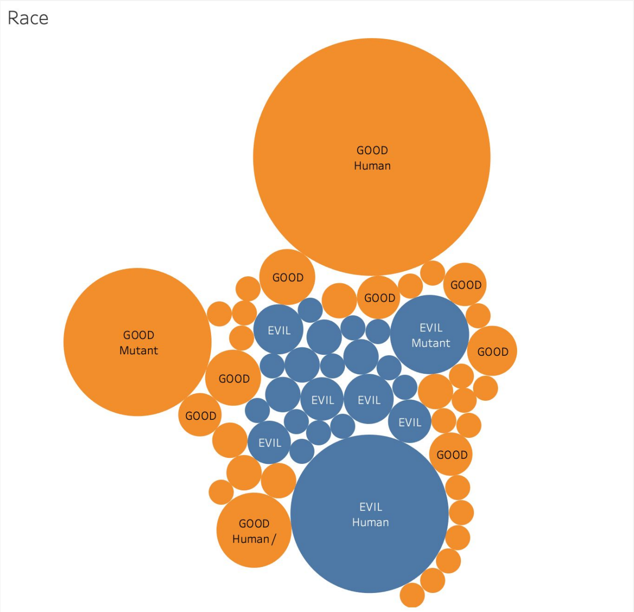
Breakdown of character alignment by gender

- Good: 243 male, 110 female
- Evil: 136 male, 28 female

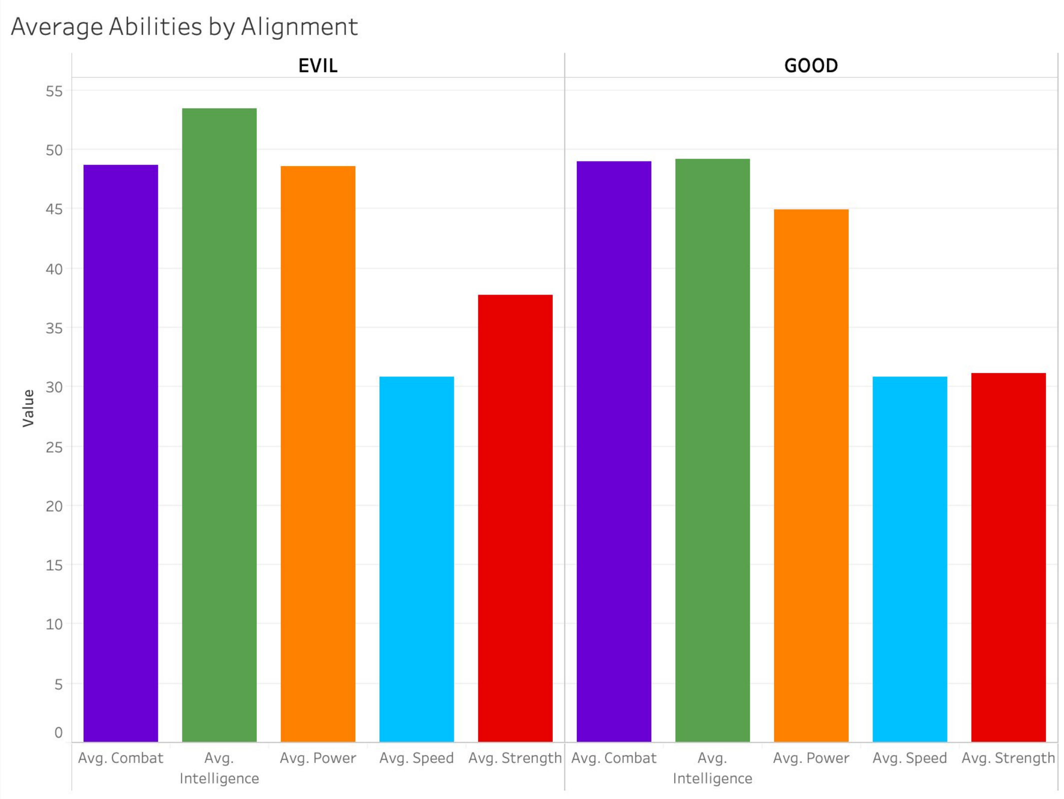
Data Analysis

- Breakdown of good and evil characters by their race
- Good chunk of data is Unknown, Human, and Mutant





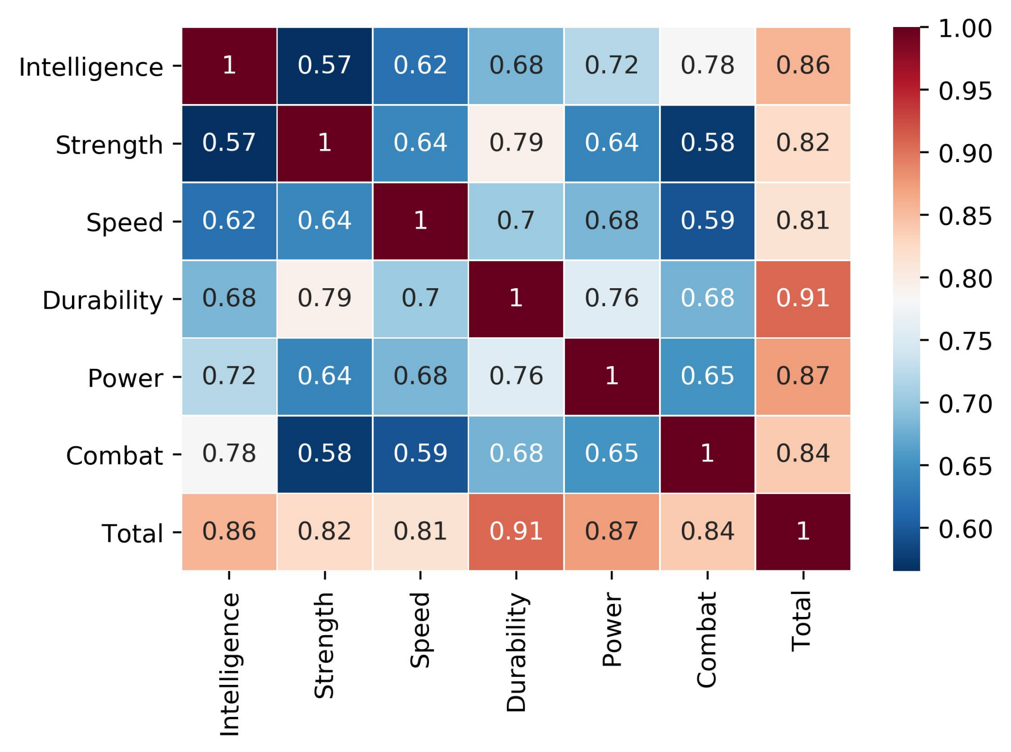
Data Analysis



- For all characters.. intelligence, power, and combat have the highest values

Data Analysis

- Pearson correlation matrix of character abilities
- Strength and Durability

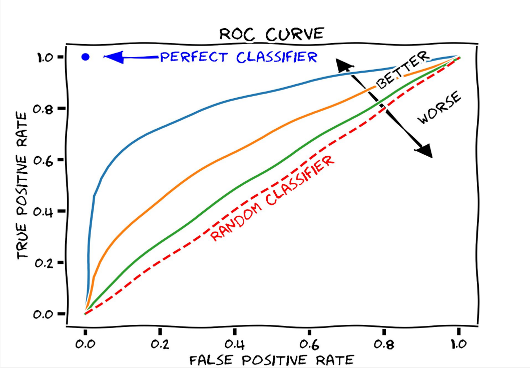


Machine Learning Models

- Classifying a character as good or evil using “Alignment” as target variable
  - Logistic Regression
  - Random Forest Classifier
  - Decision Tree
  - Support Vector Machine (SVM)
- Compare models

ML data pre-processing

- Remove highly correlated features
- Convert ‘Alignment’ and ‘Gender’ to binary variables represented as 1 and 0
- Get only true dummy values for race, hair color, eye color and of superhero abilities (agility, invisibility, web creation, and so on)
- Convert superhero matrix (strength, power, intelligence, etc..) as weights by label encoding using adaptive binning (range from 0-3 using IQRs)
- ROC (receiver operating characteristic) curve to visualize classification results



Logistic Regression Results

```
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, y_pred))
```

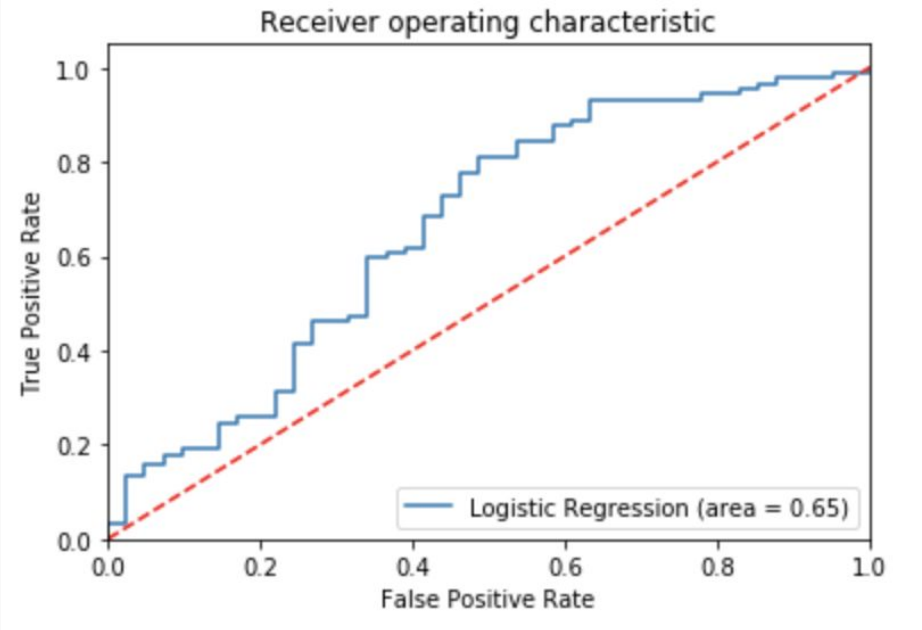
```
0.7307692307692307
```

```
matrix = confusion_matrix(y_test, y_pred)
print(matrix)
```

```
[[17 24]
 [11 78]]
```

```
report = classification_report(y_test, y_pred)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.61      0.41      0.49        41
           1       0.76      0.88      0.82        89

    accuracy                           0.73       130
   macro avg       0.69      0.65      0.65       130
weighted avg       0.72      0.73      0.71       130
```



Random Forest Classifier Results

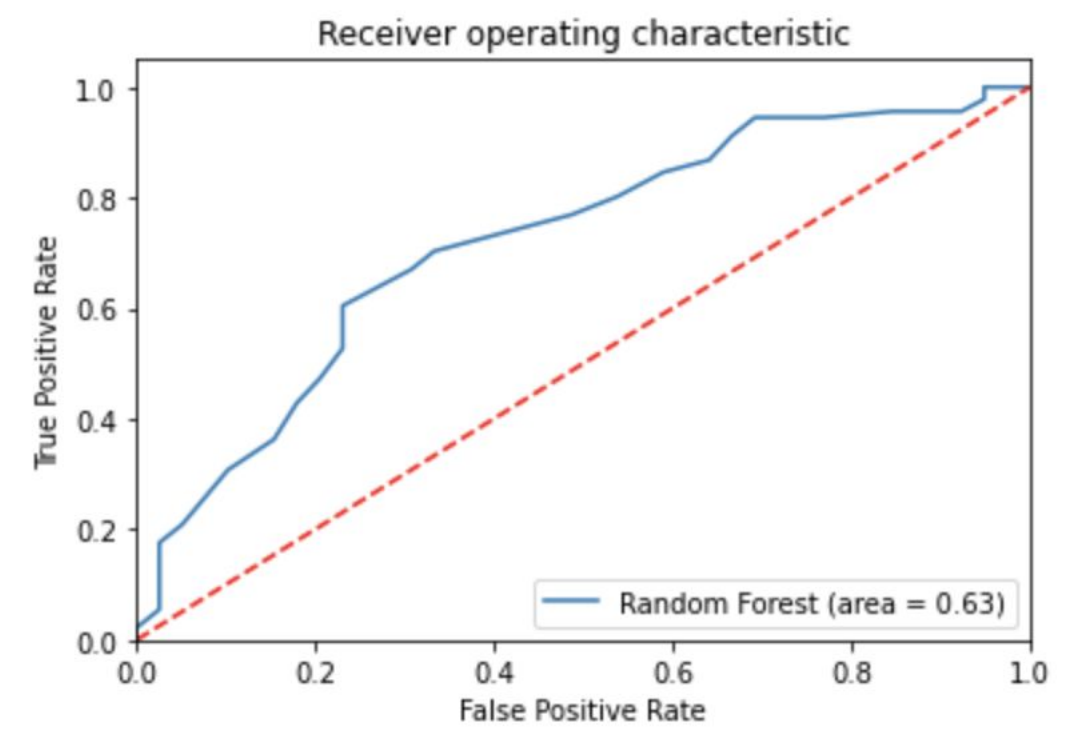
Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	12	27
Actual 1	5	86

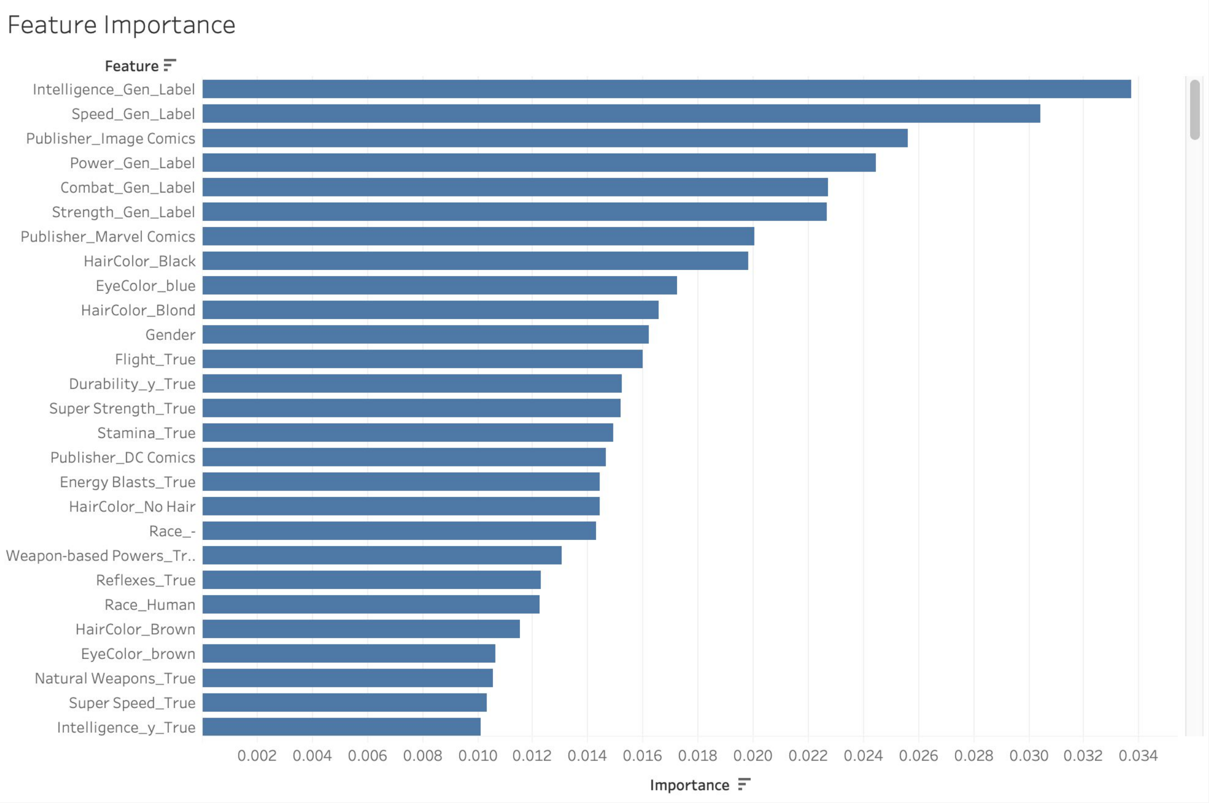
```
Accuracy Score : 0.7538461538461538
Classification Report
              precision    recall  f1-score   support

           0       0.71      0.31      0.43        39
           1       0.76      0.95      0.84        91

    accuracy                           0.75       130
   macro avg       0.73      0.63      0.64       130
weighted avg       0.74      0.75      0.72       130
```

```
# We can sort the features by their importance.
sorted(zip(rf_model.feature_importances_, X.columns), reverse=True)
```

```
[(0.033722566224920196, 'Intelligence_Gen_Label'),
 (0.03041071215605239, 'Speed_Gen_Label'),
 (0.025588502516218917, 'Publisher_Image Comics'),
 (0.024435253593627616, 'Power_Gen_Label'),
 (0.02273016367930574, 'Combat_Gen_Label'),
 (0.02265772044235813, 'Strength_Gen_Label'),
 (0.020032354416435135, 'Publisher_Marvel Comics'),
 (0.019841312393854484, 'HairColor_Black'),
 (0.017236903491103277, 'EyeColor_blue'),
 (0.016585731633076202, 'HairColor_Blond'),
 (0.016231862025821584, 'Gender'),
 (0.016000038814074413, 'Flight_True'),
 (0.015234202957567125, 'Durability_y_True'),
 (0.015206137168497837, 'Super Strength_True'),
 (0.014924949088187301, 'Stamina_True'),
 (0.014632487961048291, 'Publisher_DC Comics'),
 (0.014419925708228903, 'Energy Blasts_True'),
 (0.01441265992446499, 'HairColor_No Hair'),
```



Random Forest Classifier Results

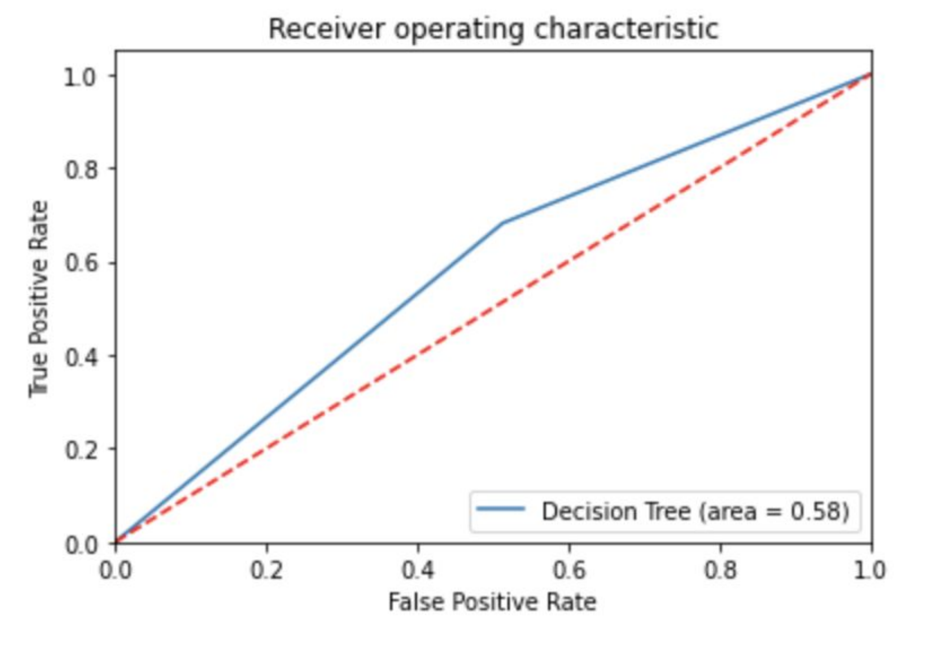


Decision Tree Results

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	20	27
Actual 1	26	83

```
Accuracy Score : 0.6602564102564102
Classification Report
              precision    recall  f1-score   support

           0       0.43      0.43      0.43        47
           1       0.75      0.76      0.76       109

    accuracy                           0.66       156
   macro avg       0.59      0.59      0.59       156
weighted avg       0.66      0.66      0.66       156
```



Support Vector Machine Results

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)
```

```
0.6384615384615384
```

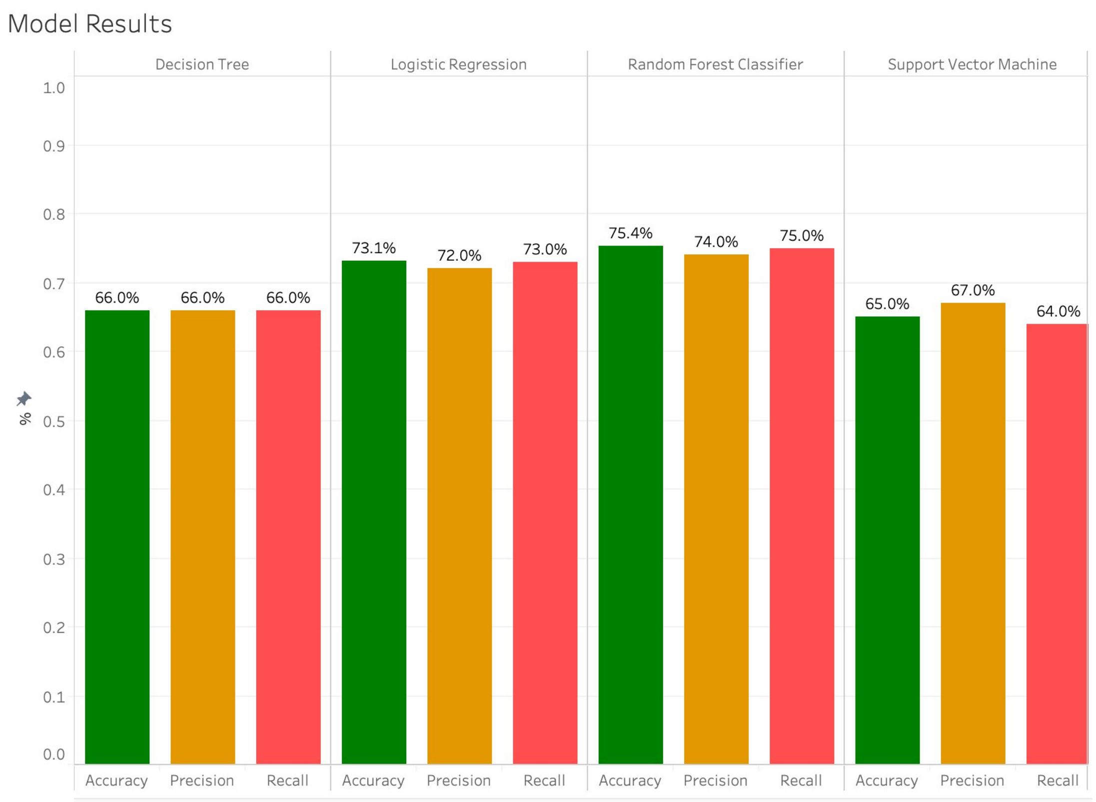
```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)
```

```
array([[24, 17],
       [30, 59]])
```

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

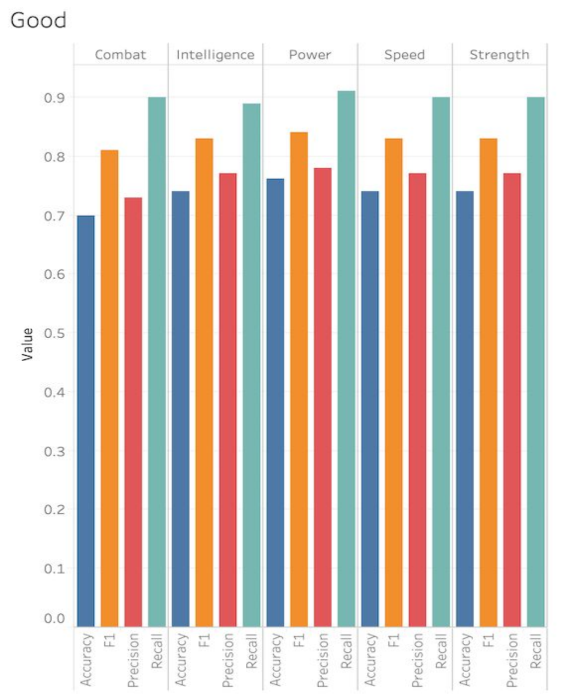
```
              precision    recall  f1-score   support

           0       0.44      0.59      0.51        41
           1       0.78      0.66      0.72        89

    accuracy                           0.64       130
   macro avg       0.61      0.62      0.61       130
weighted avg       0.67      0.64      0.65       130
```

Model Results



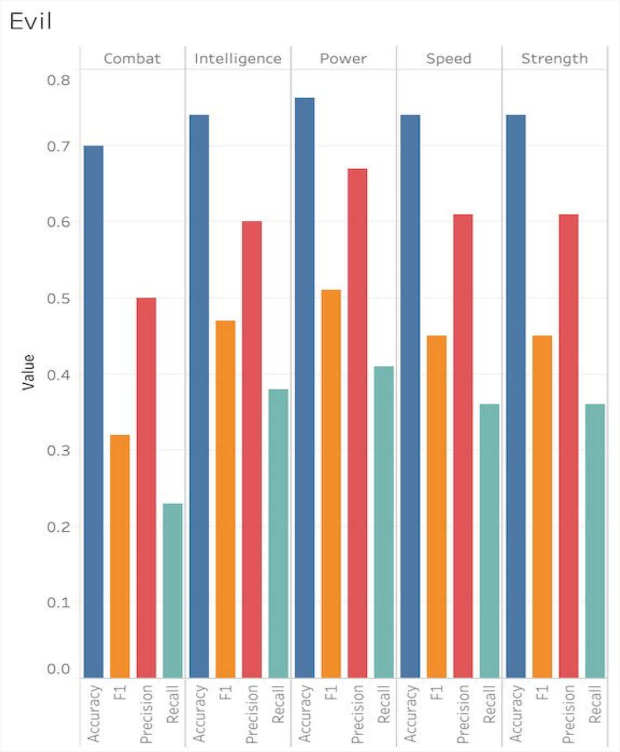
- Random forest classifier and Logistic regression were stronger models
- SVM and decision tree were weaker models

Abilities model results



- Good characters showed a well-balanced model results when predicted as good for individual abilities

Abilities model results



- Intelligence and Power showed stronger results as predictors for evil characters

Conclusion

- 75% accuracy predicting a comic book character as good or evil (somewhat strong)
- Imbalance in the dataset
- Over half are good and male characters
  - Can be explained by history of comic books’ bias towards male characters
- Evil characters are better classified based off of intelligence and/or power

Future Work

- Increase # of characters in database
- Create tiers based off of character ability stats
- Investigate similarities using k-means clustering
- Relationship between ability stats using linear regression

Tableau Dashboard

[Link to dashboard]