Machine Learning – DAT 5303 – SFMBANDD3

Assignment 2: PREDICTIVE ANALYSIS OF GAME OF THRONES CHARACTER PROSPECTIVE

Arthur Franklin Medes Filho

INTRODUCTION

The objective of this assignment is to analyze the Game of Thrones dataset, provide insightful takeaways and create machine learning model to determine the characters outcome: if they are alive or dead.

Game of Thrones is a fictional series where noble families compete for the mythical lands of Westeros, a location that resembles the Middle Ages, even though there is not a real correlation with anything in the real world, the story contains a variety of fantasy elements including dragons and zombies. The television series based on the George R. R. Martin's series of fantasy novels “A Song of Ice and Fire,” originally published in 1995 and consist of seven books. Kings, Queens, knights and lords competing in a deadly game for control of the Seven Kingdoms of Westeros, and to sit on the Iron Throne.

In an attempt to assess these causal factors of characters death, approximately 2000 observations were made from the books in a series of variables. It included if the character is a male, its culture, date of birth, the house (family), the spouse, which books (from 1 to 5) the character is present, if the parents are alive, if their heir is alive, if the spouse is alive, if the character is married, if the character is noble, the age of the character, the number of dead relatives, the popularity of the character Indicates the popularity of a character(1 = extremely popular(max), 0 = extremely unpopular(min)). These are the factors given by the case in other to determine the character outcome.

This report presents insights into factors determining if the character will live or die in the Game of Thrones series. This study did not capture other factors that could affect the character's outcome.

ANALYSIS

Before arriving at the insights drawn from this report, the dataset was explored and visualize the dataset. It is essential to comprehend the data and ensure that it is consistent, usable and of good quality. The quality of insights generated by this analysis is also highly dependent on the quality of the data, hence the data cleaning. The data cleaning required filling mistakes with correct observations and correcting misspelled words.

After the data is clean the next step was to detect and deal with outliers. Exploratory data analysis was done to summarize the factors to be analyzed. Exploration of how the houses and cultures relate to the character’s outcome lead to the creation of flags for the ones that had a greater chance of dying or remanding alive. If the house has higher chances of being dead (sum(isAlive) / count(isAlive)) is smaller than 0.5, they are flagged as a 1 in the out_house column, and the same is valid to the cultures. If all the members of the house or culture have one hundred percent chance of being alive, they were also flagged accordingly.

The visual EDA is to see what the data can tell us beyond the model. The EDA techniques were histograms and scatter plots. Additionally, a correlation analysis was done to draw insights.

INSIGHTS

Correlation analysis revealed exciting results. There was a negative correlation between if the character is alive and the date of birth, which makes sense, since many characters mentioned on the series are dead before the timeline of the story begins. For the characters born

before the start of the timeline and those in advanced age, a couple of flags were engineered to outline those who have these features.

Following that, there is also a positive relationship in if the character is male, the number or relative's dead, the popularity and their age with the character's outcome. These variables also make sense if you think of the background of the series and what is happening throughout the books: there is a war between the houses, and male characters, old enough to fight and are a prominent figure in the story are generally in a higher risk of dying.

Keeping in mind the fact that some houses and cultures are at war, a few variables were engineered to flag those specific houses and cultures that were at war and had a high probability of dying.

Using the same strategy, princes, and princesses, lords and sirs were also flagged because they are in the center of all conflict.

The qualitative variables were dropped from the analysis, and just the insightful uses of them were reflected on the final analysis, because the machine learning models do not take well the categorical variables, and using hot encoding generated too many variables, making the model to overfit the variables in the test.

In this data analysis, two different machine learning models were used: Random Forrest, which is an ensemble of a series of decision trees that predict the optimal way for the data to split into the predicted outcomes. The second model used was a Gradient Boosting Machine, which has a similar deployment, but besides generating an ensemble of decision trees, it also punishes when the machine makes a mistake.

Both methods had similar outcomes, but in the end, Random Forrest seems more insightful, giving that it can return the variable importance, it deals better with poor inputs and runs faster during a grid search.

The house and cultures were also substituted by factors, allowing the machine to use those variables in the decision-making process.

RECOMMENDATIONS

In the data analysis, some recurring factors seemed to arise. It is essential to keep the context of the books in mind when interpreting the results: there is a war where the noble houses dispute for a chance to sit in the iron throne.

That being said, the takeaways from this dataset are that houses and cultures that are more involved with the conflict have higher chances of dying, as well as central character, which tends to be more popular.

Another factor to keep in mind is that there are characters that lived before the timeline of the series began, and those who are too young to battle, therefore the year of birth is also a significant factor in the decision-making process.

My final recommendation is that the war for the iron throne should come to an end; a democracy seems to have fewer casualties.

Questions

What was your final model's highest mean AUC value after cross-validation, rounded to three decimal places?

- 0.810

EXHIBITS:

Exhibit 1

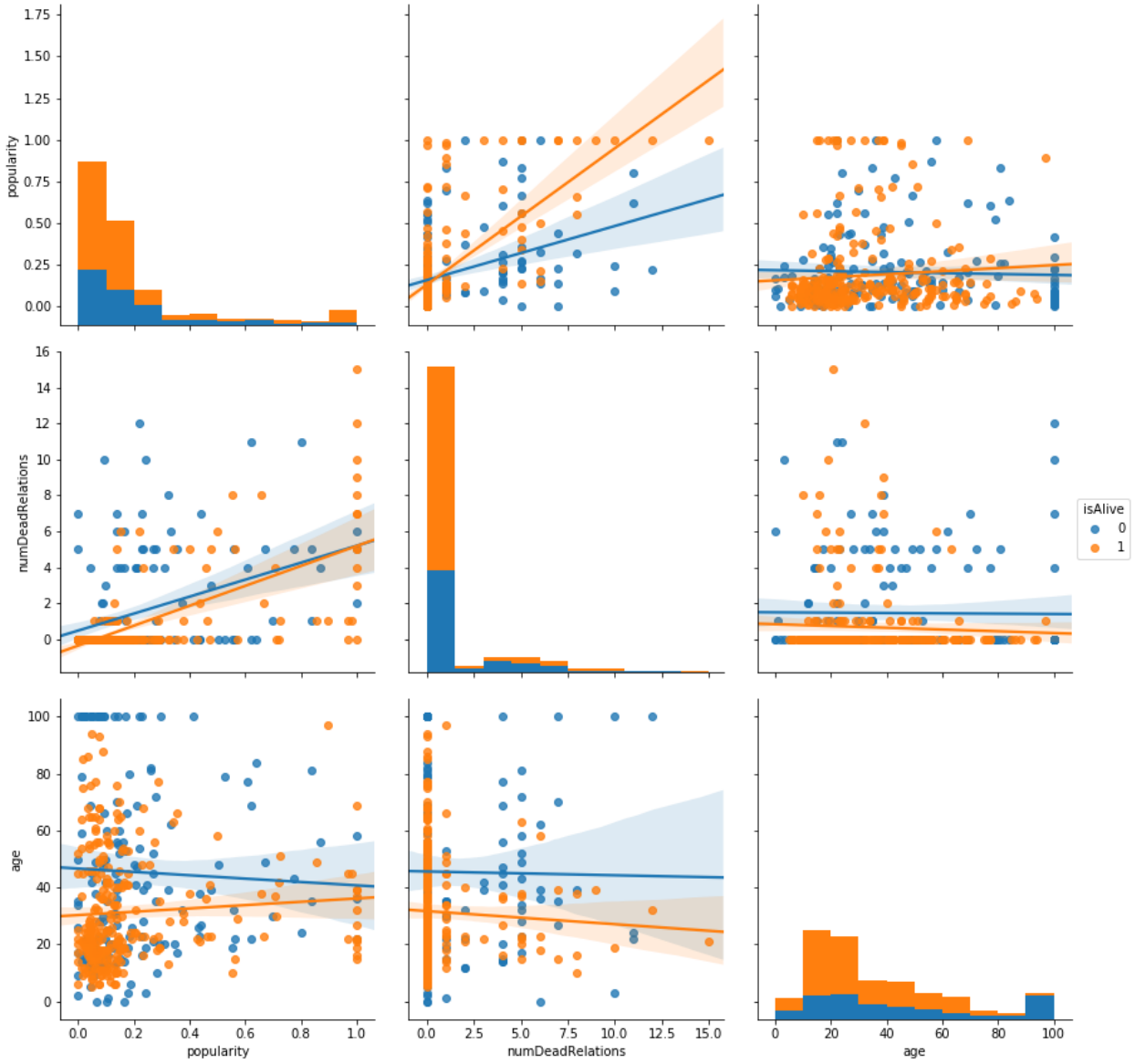


Exhibit 2

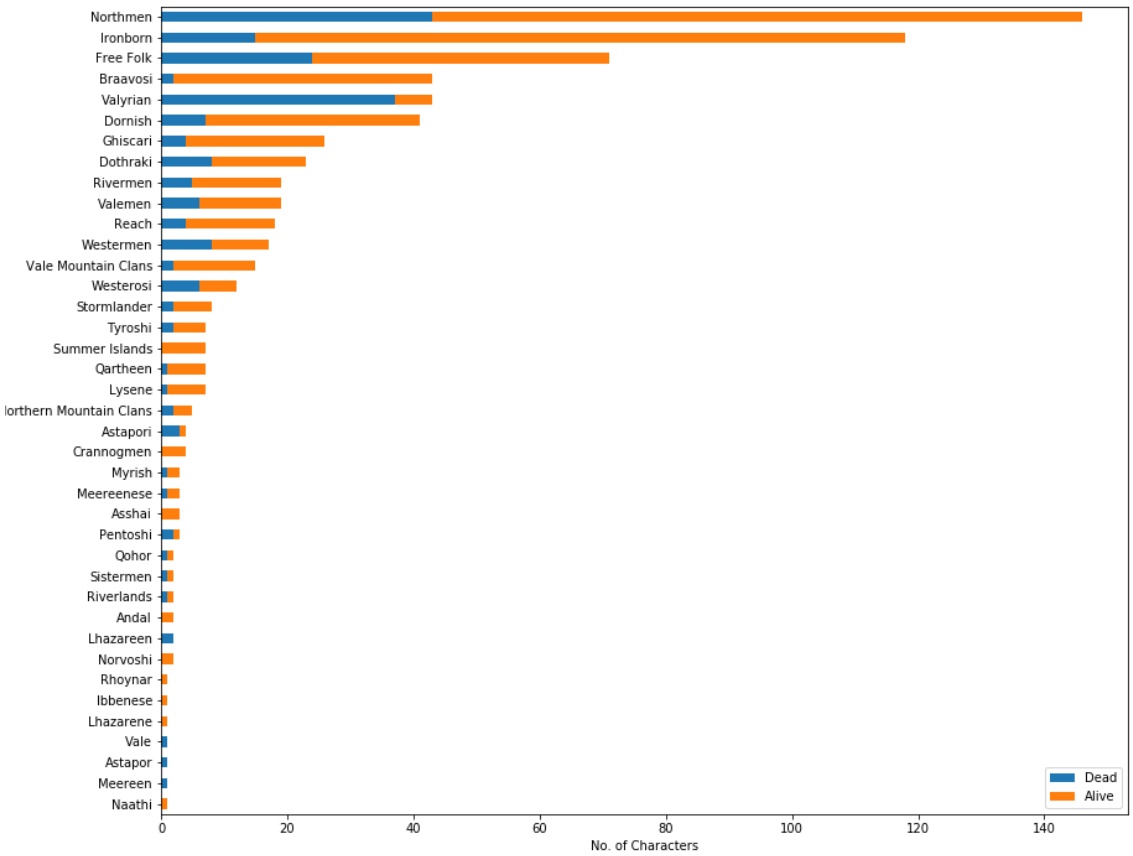

Northmen
Ironborn
Free Folk
Braavosi
Valyrian
Dornish
Ghiscari
Dothraki
Rivermen
Valemen
Reach
Westermen
Vale Mountain Clans
Westerosi
Stormlander
Tyroshi
Summer Islands
Qartheen
Lysene
lorthern Mountain Clans
Astapori
Crannogmen
Myrish
Meereenese
Asshai
Pentoshi
Qohor
Sistermen
Riverlands
Andal
Lhazareen
Norvoshi
Rhoynar
Ibbenese
Lhazarene
Vale
Astapor
Meereen
Naathi
Dead
Alive
0
20
40
60
80
100
120
140
No. of Characters

Exhibit 3

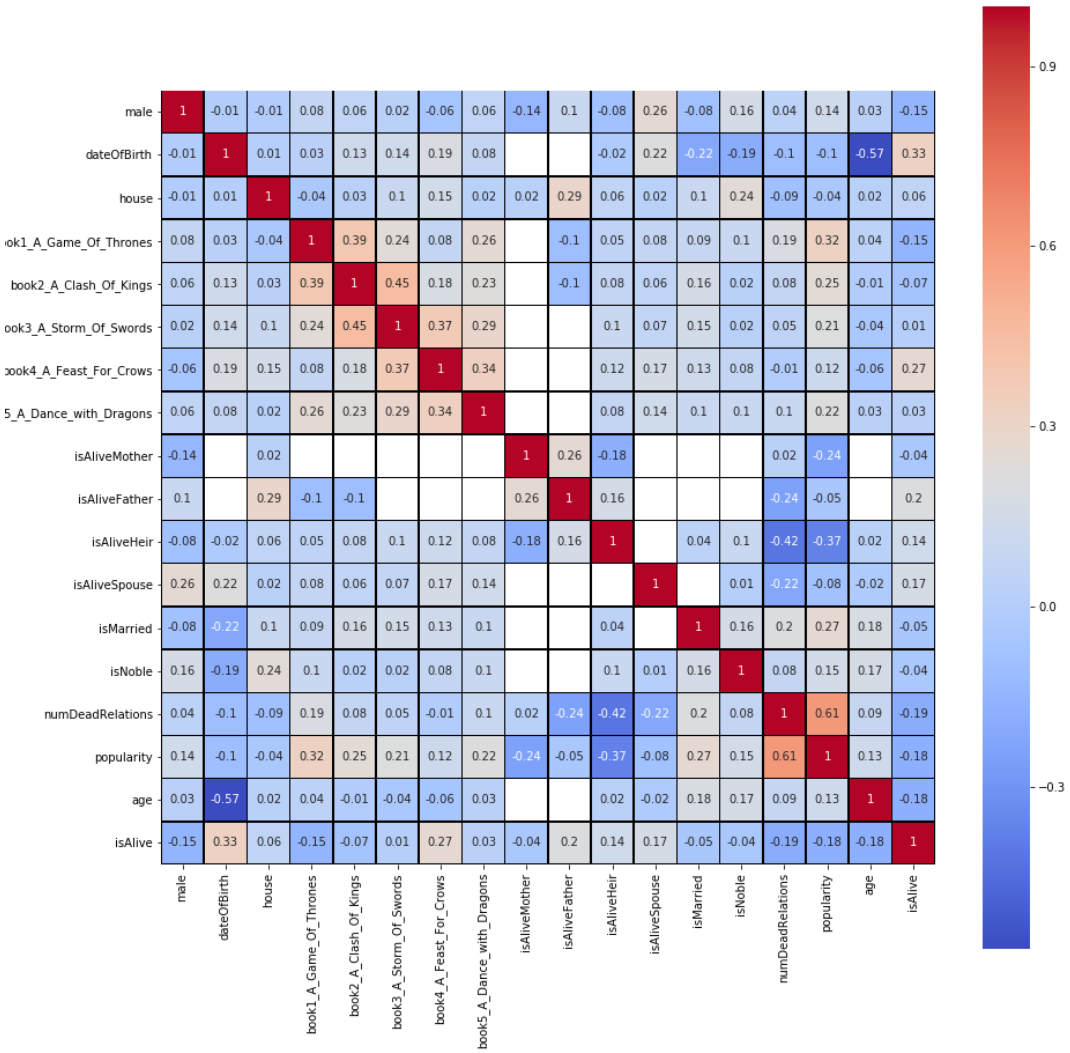

male 1 -0.01 -0.01 0.08 0.06 0.02 -0.06 0.06 -0.14 0.1 -0.08 0.26 -0.08 0.16 0.04 0.14 0.03 -0.15
dateOfBirth -0.01 1 0.01 0.03 0.13 0.14 0.19 0.08 -0.02 0.22 -0.22 -0.19 -0.1 -0.1 -0.57 0.33
house -0.01 0.01 1 -0.04 0.03 0.1 0.15 0.02 0.02 0.29 0.06 0.02 0.1 0.24 -0.09 -0.04 0.02 0.06
ook1_A_Game_Of_Thrones 0.08 0.03 -0.04 1 0.39 0.24 0.08 0.26 -0.1 0.05 0.08 0.09 0.1 0.19 0.32 0.04 -0.15
book2_A_Clash_Of_Kings 0.06 0.13 0.03 0.39 1 0.45 0.18 0.23 -0.1 0.08 0.06 0.16 0.02 0.08 0.25 -0.01 -0.07
ook3_A_Storm_Of_Swords 0.02 0.14 0.1 0.24 0.45 1 0.37 0.29 0.1 0.07 0.15 0.02 0.05 0.21 -0.04 0.01
ook4_A_Feast_For_Crows -0.06 0.19 0.15 0.08 0.18 0.37 1 0.34 0.12 0.17 0.13 0.08 -0.01 0.12 -0.06 0.27
5_A_Dance_with_Dragons 0.06 0.08 0.02 0.26 0.23 0.29 0.34 1 0.08 0.14 0.1 0.1 0.1 0.22 0.03 0.03
isAliveMother -0.14 0.02 1 0.26 -0.18 0.02 -0.24 -0.04
isAliveFather 0.1 0.29 -0.1 -0.1 0.26 1 0.16 -0.24 -0.05 0.2
isAliveHeir -0.08 -0.02 0.06 0.05 0.08 0.1 0.12 0.08 -0.18 0.16 1 0.04 0.1 -0.42 -0.37 0.02 0.14
isAliveSpouse 0.26 0.22 0.02 0.08 0.06 0.07 0.17 0.14 1 0.01 -0.22 -0.08 -0.02 0.17
isMarried -0.08 -0.22 0.1 0.09 0.16 0.15 0.13 0.1 0.04 1 0.16 0.2 0.27 0.18 -0.05
isNoble 0.16 -0.19 0.24 0.1 0.02 0.02 0.08 0.1 0.1 0.01 0.16 1 0.08 0.15 0.17 -0.04
numDeadRelations 0.04 -0.1 -0.09 0.19 0.08 0.05 -0.01 0.1 0.02 -0.24 -0.42 -0.22 0.2 0.08 1 0.61 0.09 -0.19
popularity 0.14 -0.1 -0.04 0.32 0.25 0.21 0.12 0.22 -0.24 -0.05 -0.37 -0.08 0.27 0.15 0.61 1 0.13 -0.18
age 0.03 -0.57 0.02 0.04 -0.01 -0.04 -0.06 0.03 0.02 -0.02 0.18 0.17 0.09 0.13 1 -0.18
isAlive -0.15 0.33 0.06 -0.15 -0.07 0.01 0.27 0.03 -0.04 0.2 0.14 0.17 -0.05 -0.04 -0.19 -0.18 -0.18 1
male
dateOfBirth
house
book1_A_Game_Of_Thrones
book2_A_Clash_Of_Kings
book3_A_Storm_Of_Swords
book4_A_Feast_For_Crows
book5_A_Dance_with_Dragons
isAliveMother
isAliveFather
isAliveHeir
isAliveSpouse
isMarried
isNoble
numDeadRelations
popularity
age
isAlive
0.9
0.6
0.3
0.0
−0.3

Exhibit 4

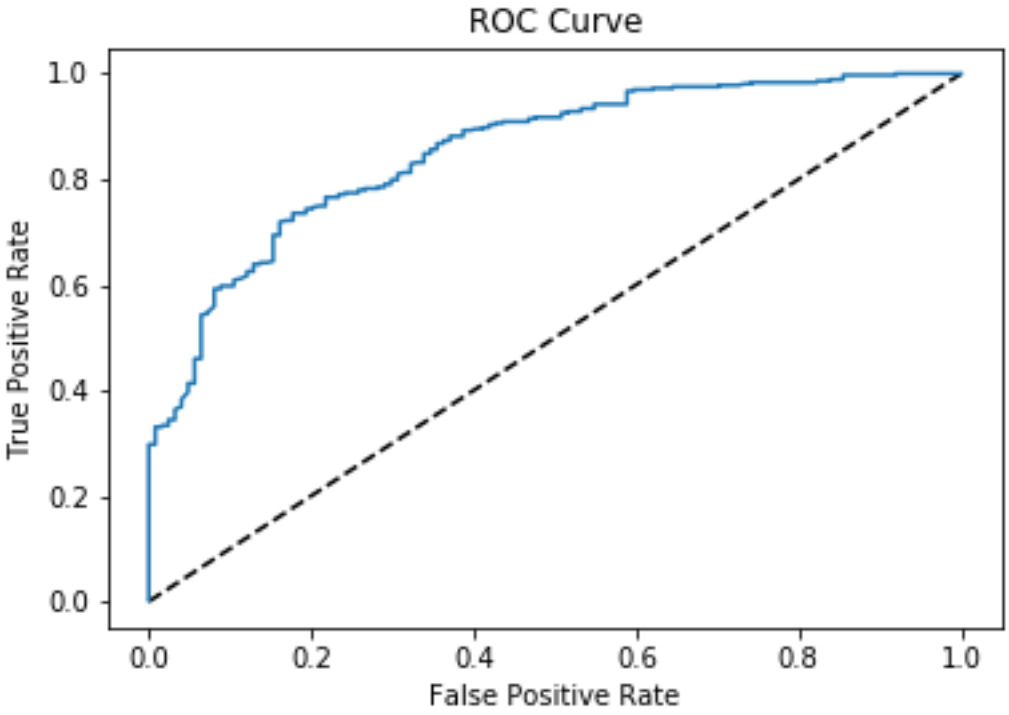
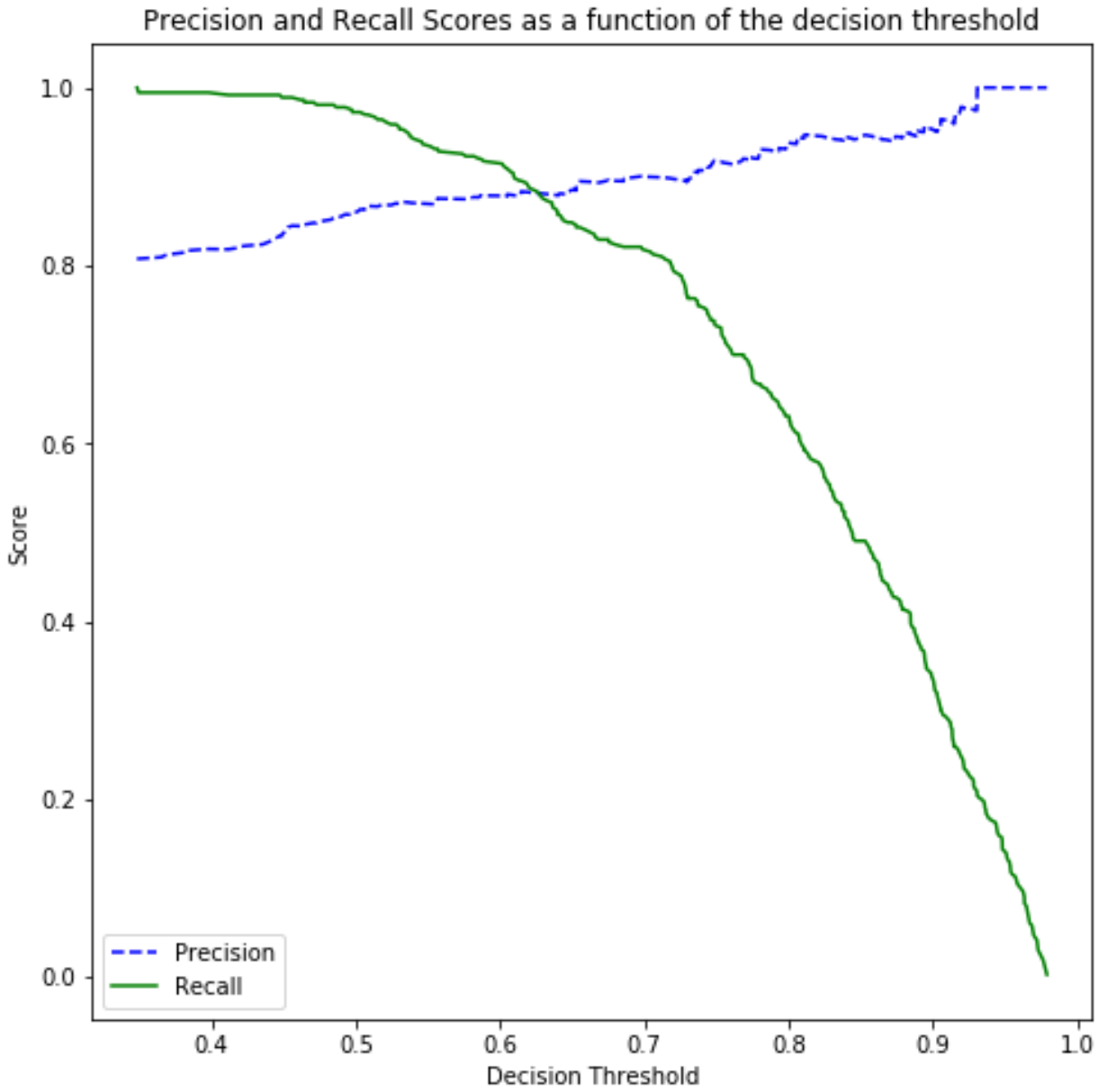

ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Exhibit 5



REFERENCE:

Game of Thrones. (2011, April 17). Retrieved from https://www.imdb.com/title/tt0944947/

What Is Game of Thrones? (n.d.). Retrieved from https://www.dummies.com/art-center/performing-arts/filmmaking/what-is-game-of-thrones/

Sklearn.calibration.CalibratedClassifierCV¶. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html#sklearn.calibration.CalibratedClassifierCV

Mendes, A. (2019, March 16). ArthurFMendes/GOT_prediction. Retrieved from https://github.com/ArthurFMendes/GOT_prediction

Glossary of Common Terms and API Elements¶. (n.d.). Retrieved from https://scikit-learn.org/stable/glossary

Ferreira, H., & Ferreira, H. (2018, April 04). Confusion matrix and other metrics in machine learning. Retrieved from https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning-894688cb1c0a