HULT
INTERNATIONAL
BUSINESS SCHOOL

Assignment 2: Written Assessment – Classification Modelling Case Study

ANALYSIS WRITE UP

Machine Learning - DAT-5303 - SFMBANDD1
Professor: Chase Kusterer

March 18th, 2019

Anant Kumar

Table of Contents

Introduction

The dataset for Game of Thrones (GoT) available for our assessment is 26 columns carrying qualitative, numerical and factor variables. Many of these variables are having missing values to the tune of 99% and also some mis-spelled words have been found. The assignment is to predict data for column 'isAlive' which is a binary data and hence a problem of logistic regression.

Key Insights

The dataset of GoT was scrubbed and found some of the followings are highly unsuitable for analysis. Before start of any kind of imputation, flags for missing values and outliers were created.

Title: 52% missing value with 263 unique values in the column which is 28% of the available record for the column. If we make it average, we can say that no value has more than four repition in the column, Though, in actual, some of the value is highly repetitive but most of these are so sparse. We need to fill in the missing values and also lower the variable count in the column. To achieve these, I treated the missing value as "Unknown" and all those sparse entries (anything appearing less than 15 times) as "Other" to make the column suitable for analysis.

Culture: 65% missing values and just 9% unique values. Treated the same way as above.

dateOfBirth: Since a column is available for age of the characters, I preferred to discard the data.

mother: 99% missing values, 85% unique records from the available set of data. With such a high number of missing value, I found the data unsuitable for any kind of analysis.

father: 99% missing values and 80% unique records. Discarded the data for inclusion in analysis for the reasons.

heir: Discarded same as above for the same reasons. Missing values – 98%

spouse: It is name with 92% unique values. Preferred not to use in the analysis.

isAliveMother, isAliveFather, isAliveHeir, isAliveSpouse: all are binary data and can not be imputed. After creating missing flag.

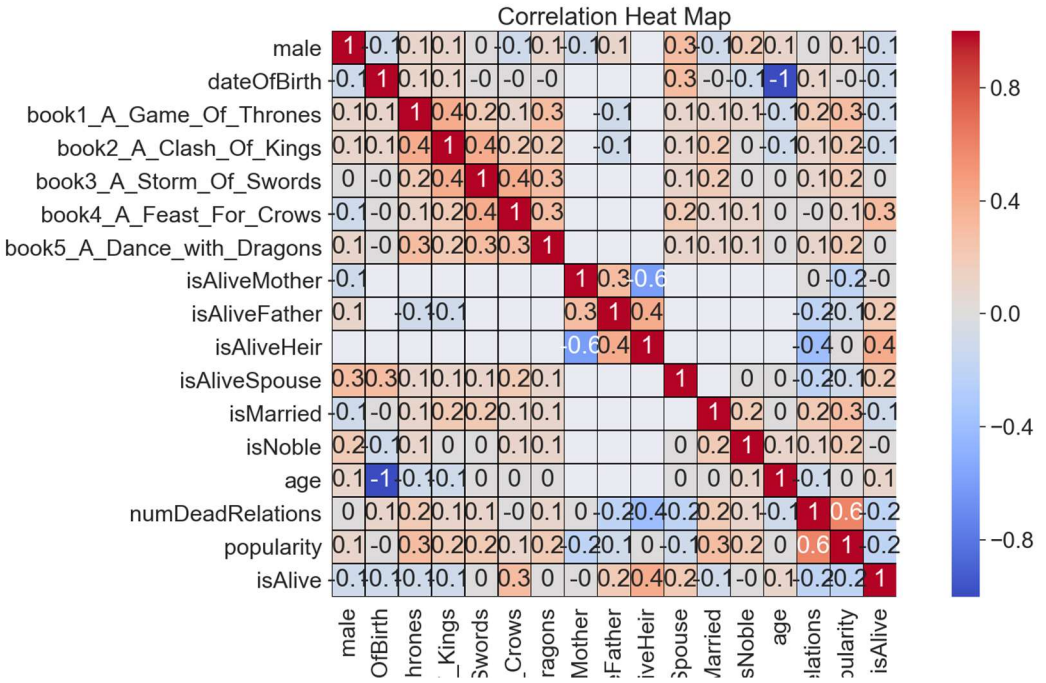
age: Age has 77% missing values. It might prove to be an important variable and needs to imputed. To impute, going with mean or median age of the available dataset was one way. But, in my analysis, I observed, the median age of characters appearing first in different volumes of the book are different. Median age for character appearing first in first book, 2nd book, 3rd book, 4th book, 5th book and probably 6th (all those characters who did not appear in any of the first 5 books) book is 34, 21, 21, 27, 55 and 45 in that order. So while imputing data for age, I considered which book the character appeared first.

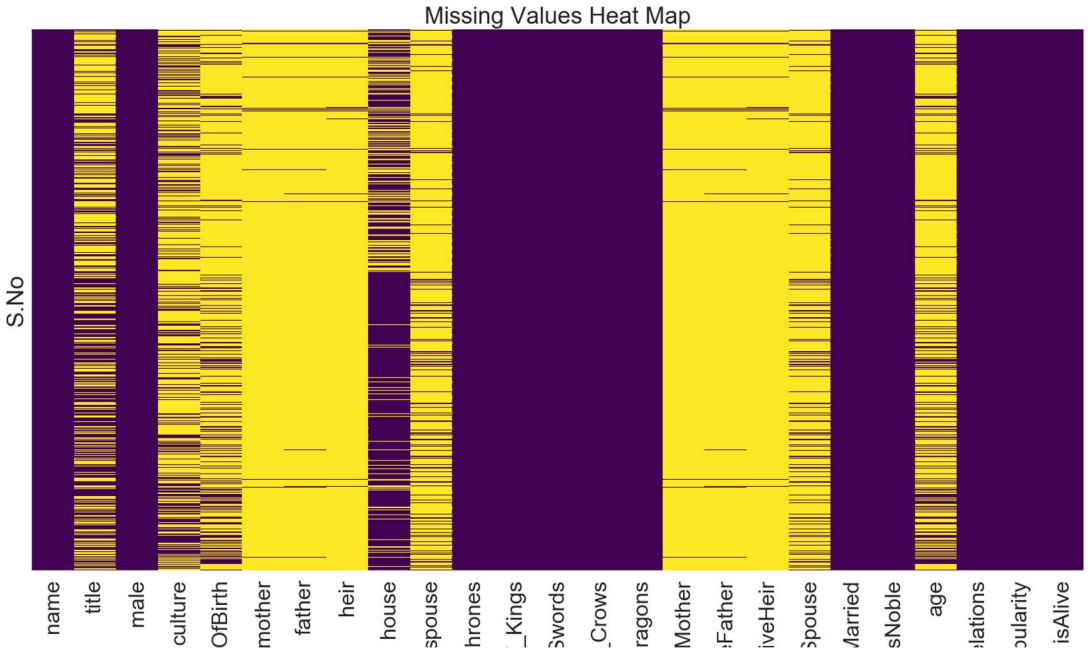
After completion of data cleaning, data were split (0.9: 0.1 train-test split) for training and testing of models. Models were initiated and fitted for the Logistic regression, KNN, Decision Tree, Random Forest, Support vector classifier and Gradient Boosting classifier. The model average score after CV = 3 are 0.777, 0.766, 0.722, 0.788, 0.766 and 0.739 respectively.

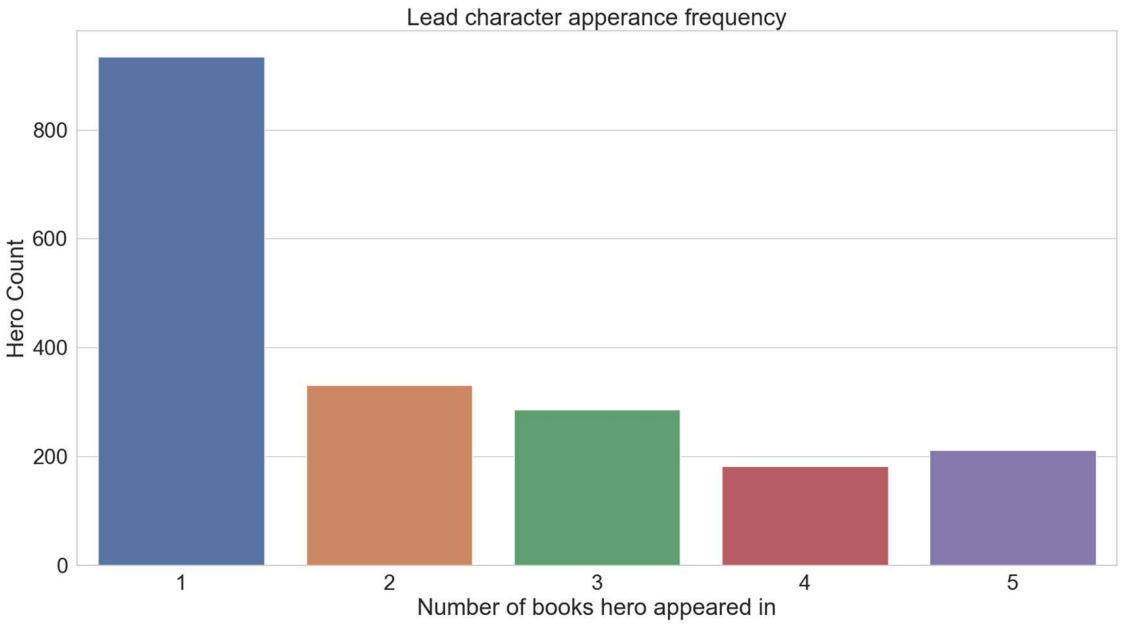
Conclusions

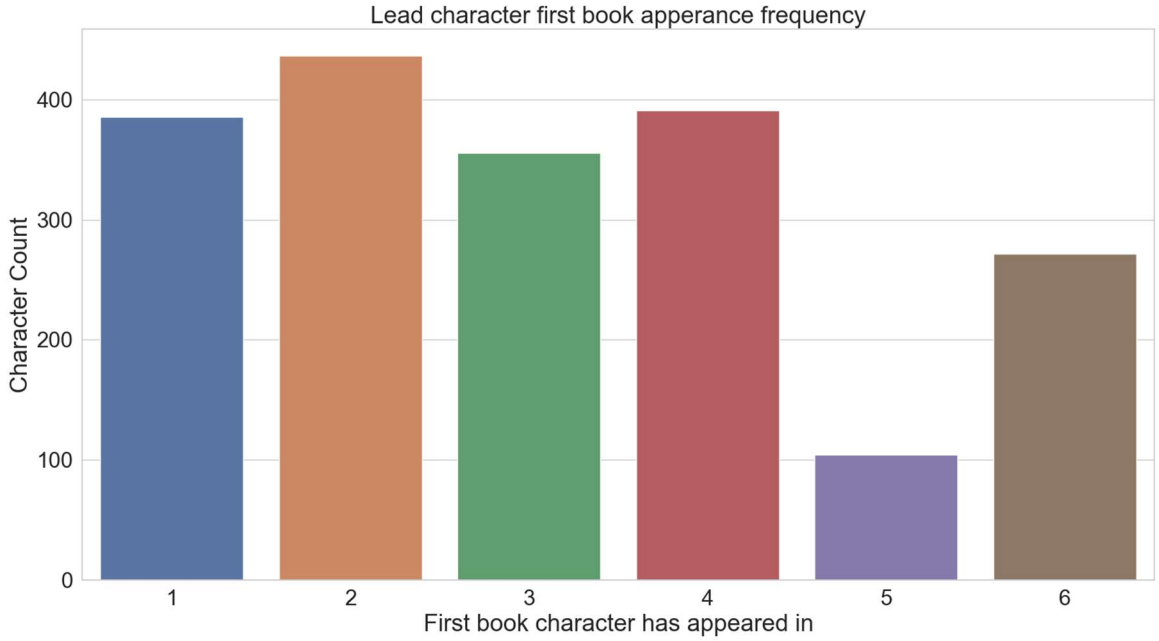
The result obtained from the above model building are consistent with accuracy around 0.78. The will likely to improve with more future engineering.

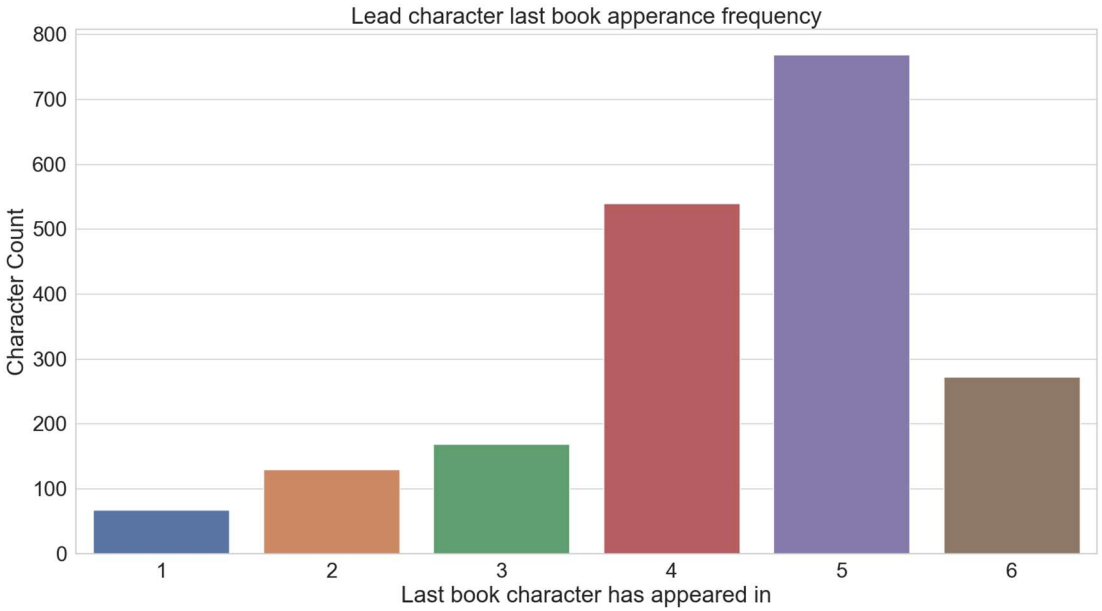
Appendix

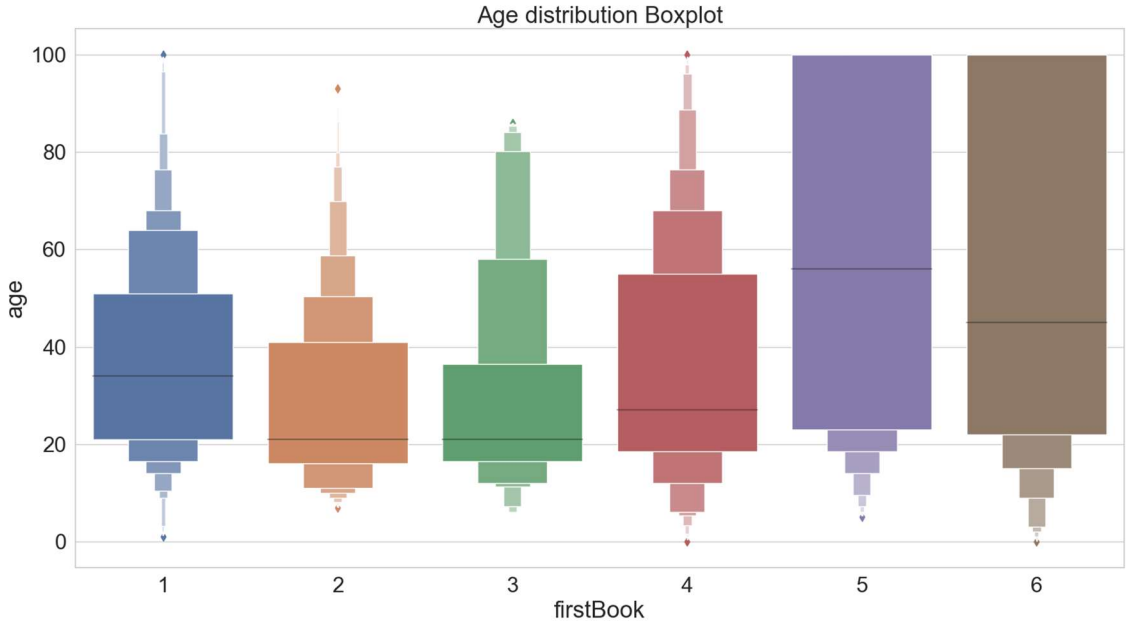

Correlation Heat Map
male
dateOfBirth
book1_A_Game_Of_Thrones
book2_A_Clash_Of_Kings
book3_A_Storm_Of_Swords
book4_A_Feast_For_Crows
book5_A_Dance_with_Dragons
isAliveMother
isAliveFather
isAliveHeir
isAliveSpouse
isMarried
isNoble
age
numDeadRelations
popularity
isAlive
0.8
0.4
0.0
−0.4
−0.8

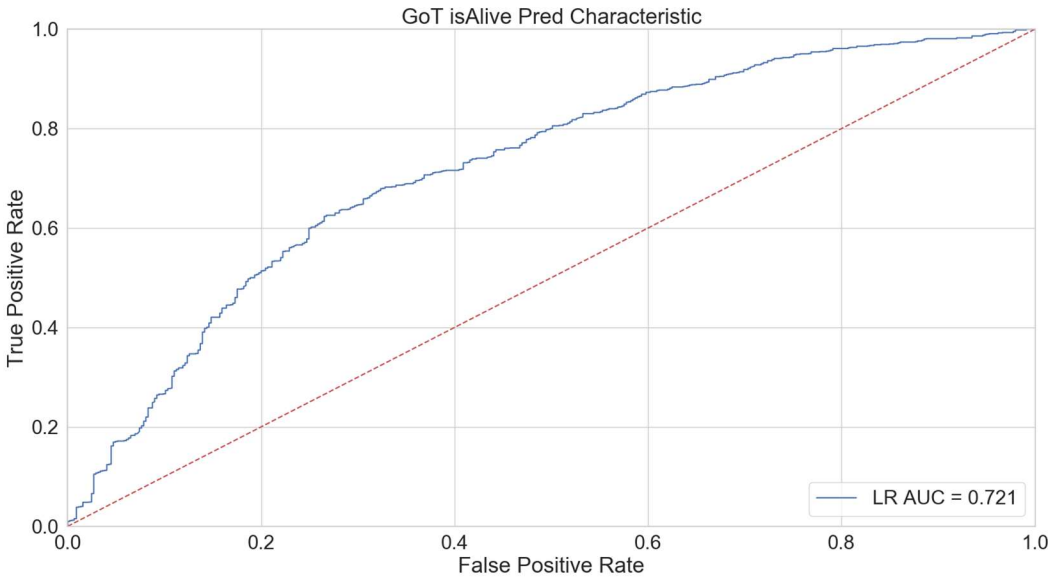

Missing Values Heat Map
S.No
name
title
male
culture
OfBirth
mother
father
heir
house
spouse
hrones
_Kings
Swords
_Crows
ragons
Mother
Father
iveHeir
Spouse
Married
sNoble
age
lations
pularity
isAlive

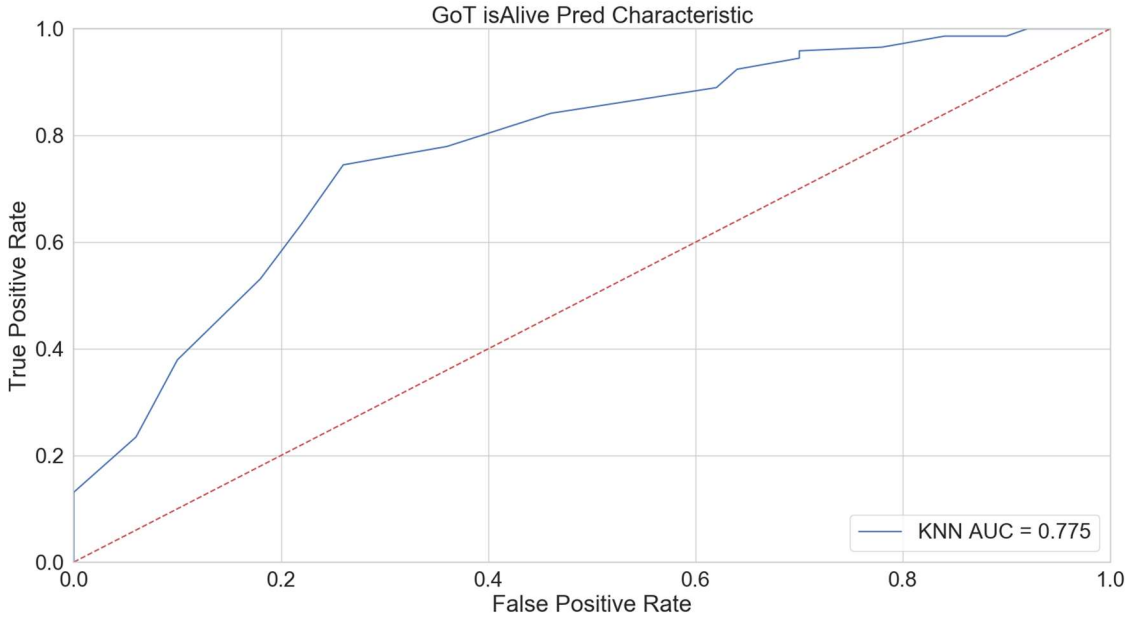

Lead character apperance frequency
Hero Count
800
600
400
200
0
1
2
3
4
5
Number of books hero appeared in

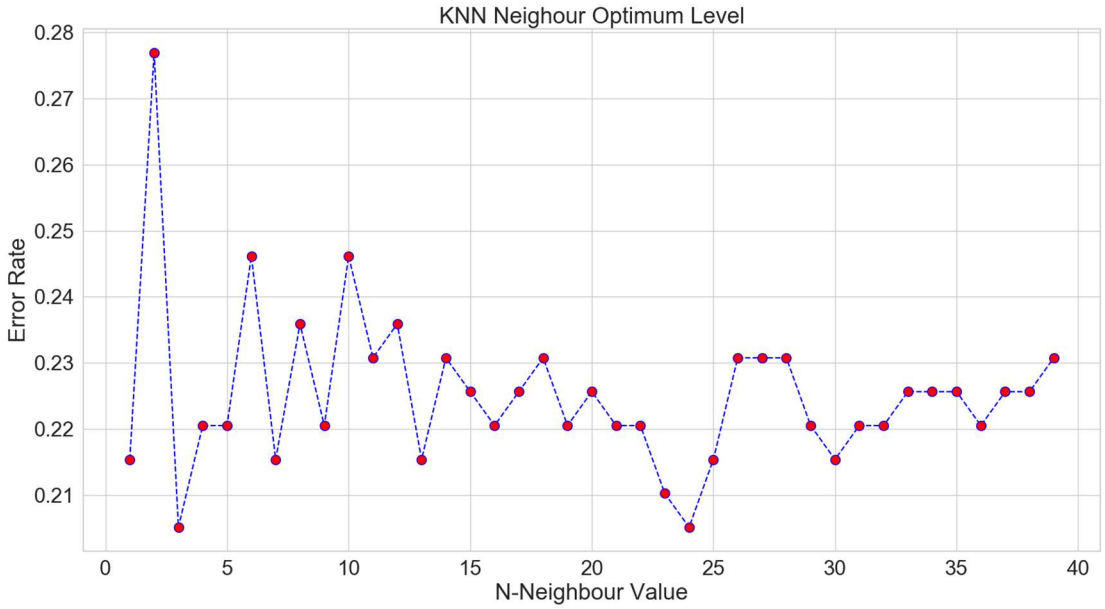

Lead character first book apperance frequency
Character Count
400
300
200
100
0
1
2
3
4
5
6
First book character has appeared in

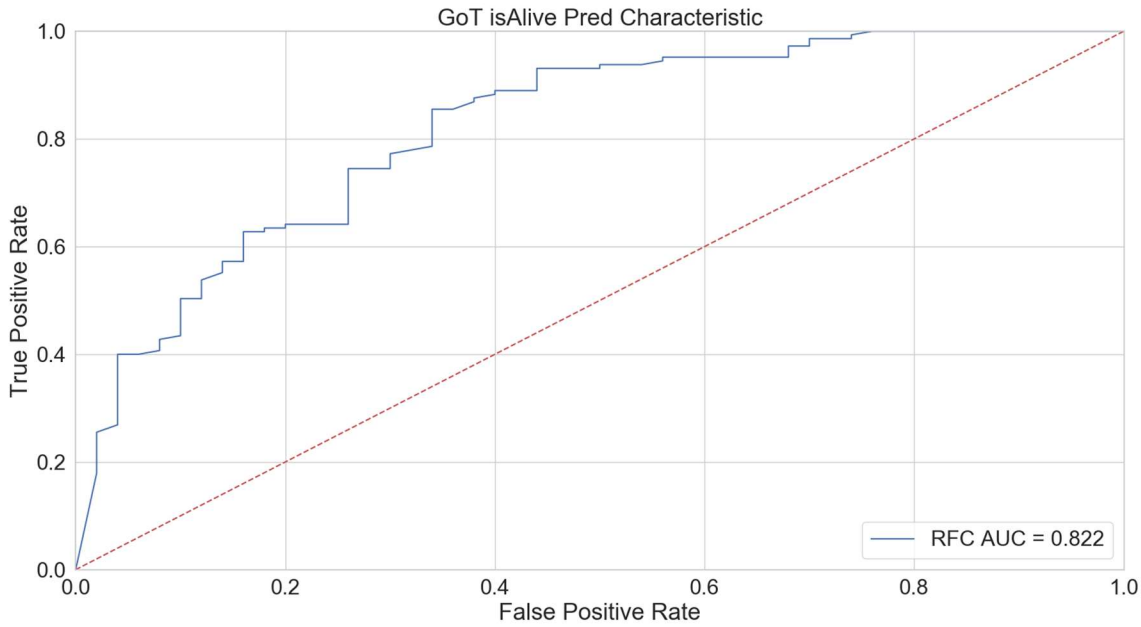

Lead character last book apperance frequency
Character Count
800
700
600
500
400
300
200
100
0
1
2
3
4
5
6
Last book character has appeared in


Age distribution Boxplot
age
100
80
60
40
20
0
1
2
3
4
5
6
firstBook


GoT isAlive Pred Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LR AUC = 0.721


GoT isAlive Pred Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
KNN AUC = 0.775


KNN Neighour Optimum Level
Error Rate
N-Neighbour Value
0.28
0.27
0.26
0.25
0.24
0.23
0.22
0.21
0
5
10
15
20
25
30
35
40


GoT isAlive Pred Characteristic
True Positive Rate
False Positive Rate
RFC AUC = 0.822
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0