GE Credit/Financing Risk Data Model Results

Random Forest Model

Charles Adkins

Southern New Hampshire University

What do we currently know?

Dataset =

700 No Defaults & 300 Defaults

- 30% error rate

Goal =

minimize Default events by lowering error rate of application approvals

How do we approach the issue?

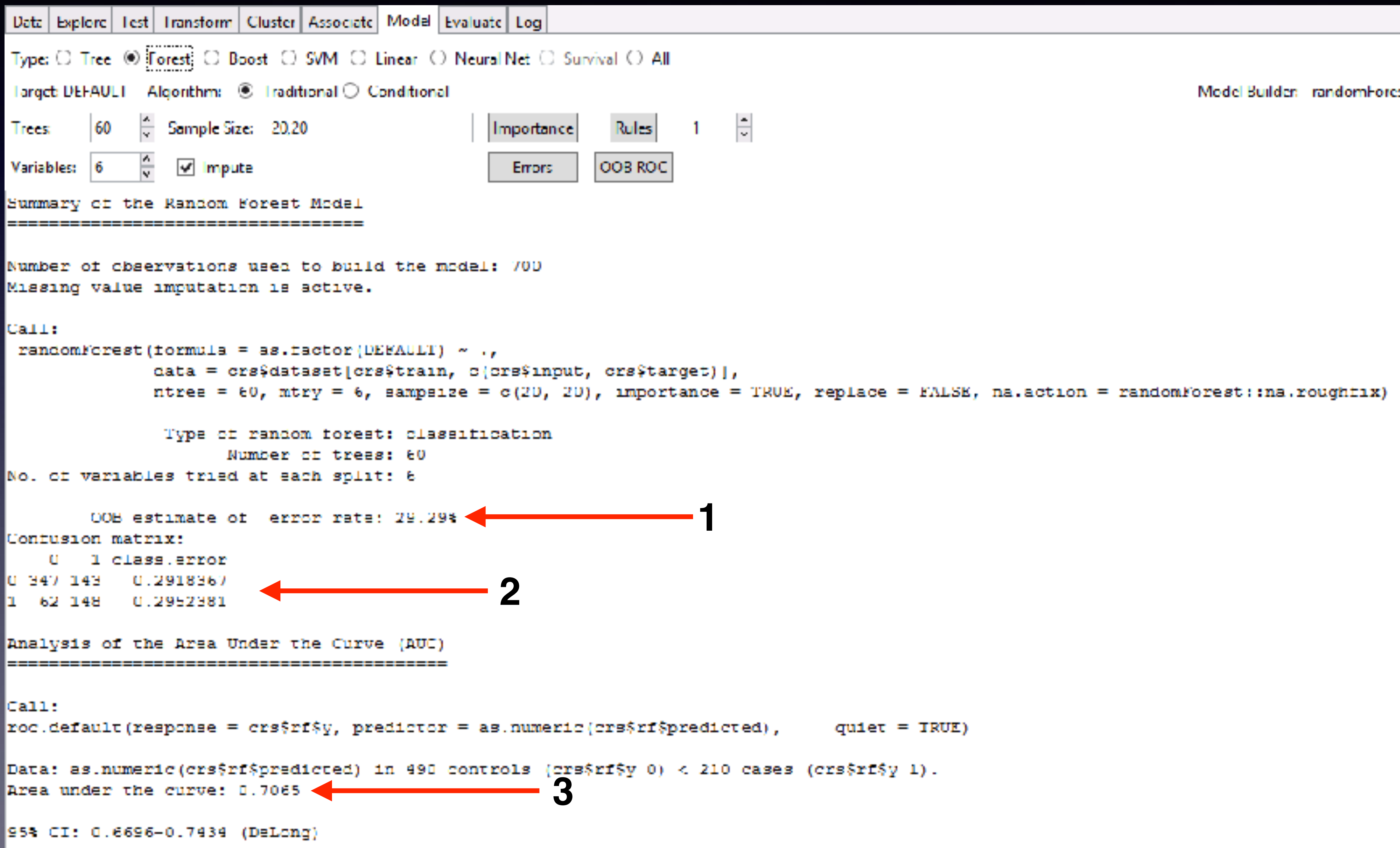
Utilizing the Random Forest model will provide us with reliable results by accounting for data bias or inaccurate influence and by identifying important variables.

By identifying the important variables, the model will establish criteria that is able to predict whether an applicant will default or not, based upon the values of those variables.

The prediction will assist the credit branches in the application decision making process by providing them with numerical data that supports the likelihood of a certain event, Default or No Default.

Random Forest Model Results

Summary



Which variables are important?

- Important variables establish which variables are most influential in determining the outcome of Default or No Default

- By identifying those variables, we can more quickly identify the presence of these variables and make a determination of application approval or disapproval

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
CHK_ACCT	5.29	5.72	5.87	2.41
HISTORY	2.86	3.17	3.50	1.65
DURATION	2.12	2.53	3.03	1.48
EMPLOYMENT	2.31	2.50	3.02	1.38
NEW_CAR	1.71	2.41	2.85	0.44
SAV_ACCT	1.60	3.40	2.85	1.26
PROP_UNKN_NONE	2.06	1.30	2.59	0.53
MALE_SINGLE	2.39	-0.59	2.34	0.25
RADIO.TV	1.99	1.71	2.22	0.21
GUARANTOR	2.03	1.96	2.07	0.15
AGE	1.55	0.44	1.70	2.31
MALE_DIV	1.44	1.01	1.47	0.11
AMOUNT	0.34	1.77	1.28	2.02
CO.APPLICANT	1.61	0.49	1.22	0.18
NUM_DEPENDENTS	0.16	2.22	1.20	0.21
OTHER_INSTALL	1.39	0.35	1.16	0.47
FOREIGN	0.54	0.86	1.11	0.09
RENT	0.78	-0.16	1.03	0.22
RETRAINING	0.77	0.54	0.88	0.19
REAL_ESTATE	0.62	-0.05	0.75	0.43
NUM_CREDITS	0.91	-0.41	0.70	0.20
PRESENT_RESIDENT	0.62	-0.26	0.55	0.84
JOB	0.30	0.07	0.34	0.51
OWN_RES	0.54	-0.75	0.29	0.37
EDUCATION	-0.45	0.76	0.26	0.10
INSTALL_RATE	0.49	-0.69	0.13	0.77
MALE_MAR_or_WID	-0.21	0.63	0.00	0.33
USED_CAR	-0.09	-0.23	-0.18	0.26
FURNITURE	-0.57	0.12	-0.54	0.24
TELEPHONE	-0.91	-0.07	-1.38	0.29

Model Error

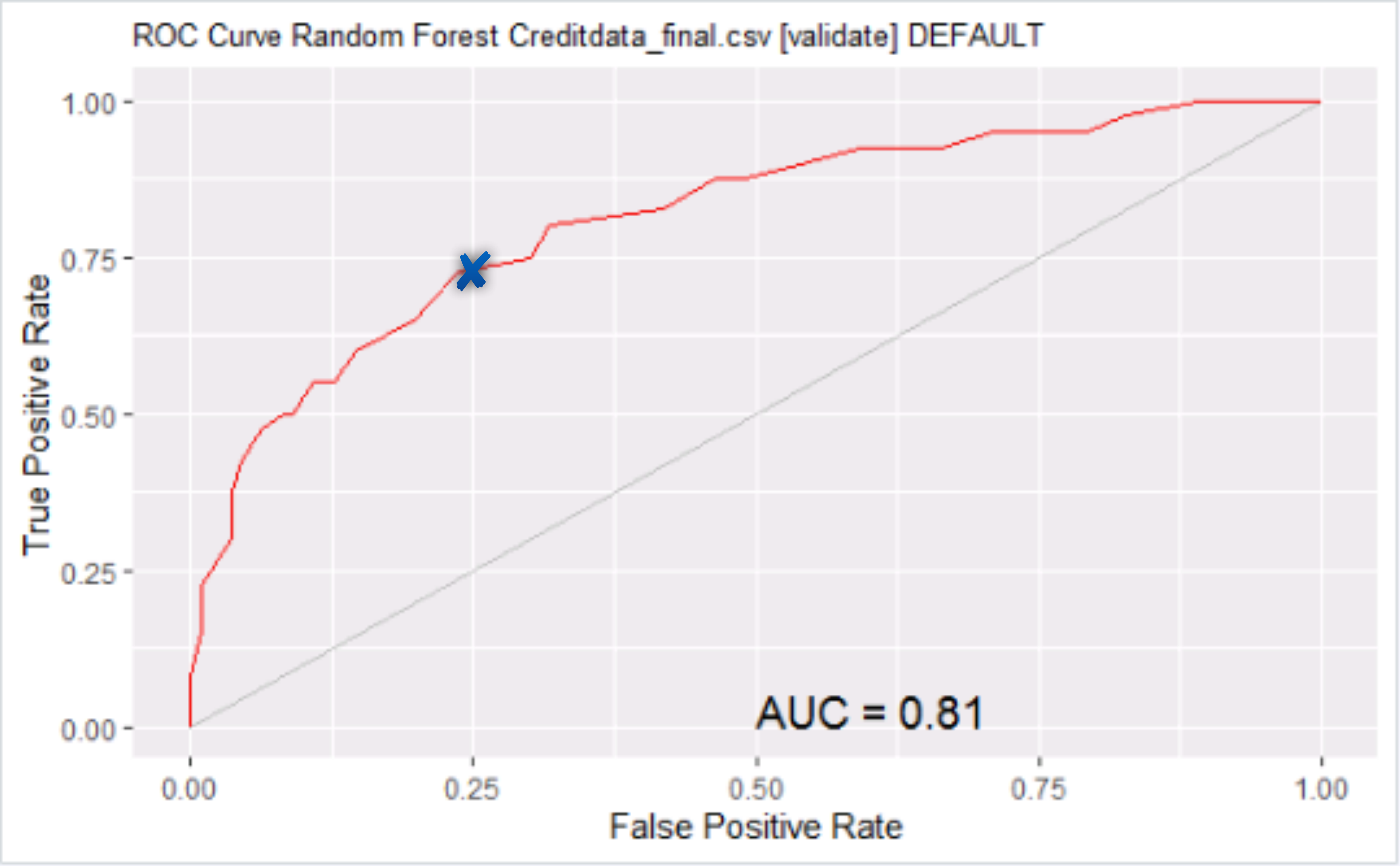
```
Error matrix for the Random Forest model on Creditdata_final.csv [test] (counts):

      Predicted
Actual  0  1 Error
     0 81 19    19
     1 14 36    28

Error matrix for the Random Forest model on Creditdata_final.csv [test] (proportions):

      Predicted
Actual    0    1 Error
     0 51.0 12.7    19
     1  9.3 21.0    28

Overall error: 22%, Averaged class error: 23.5%
```

1

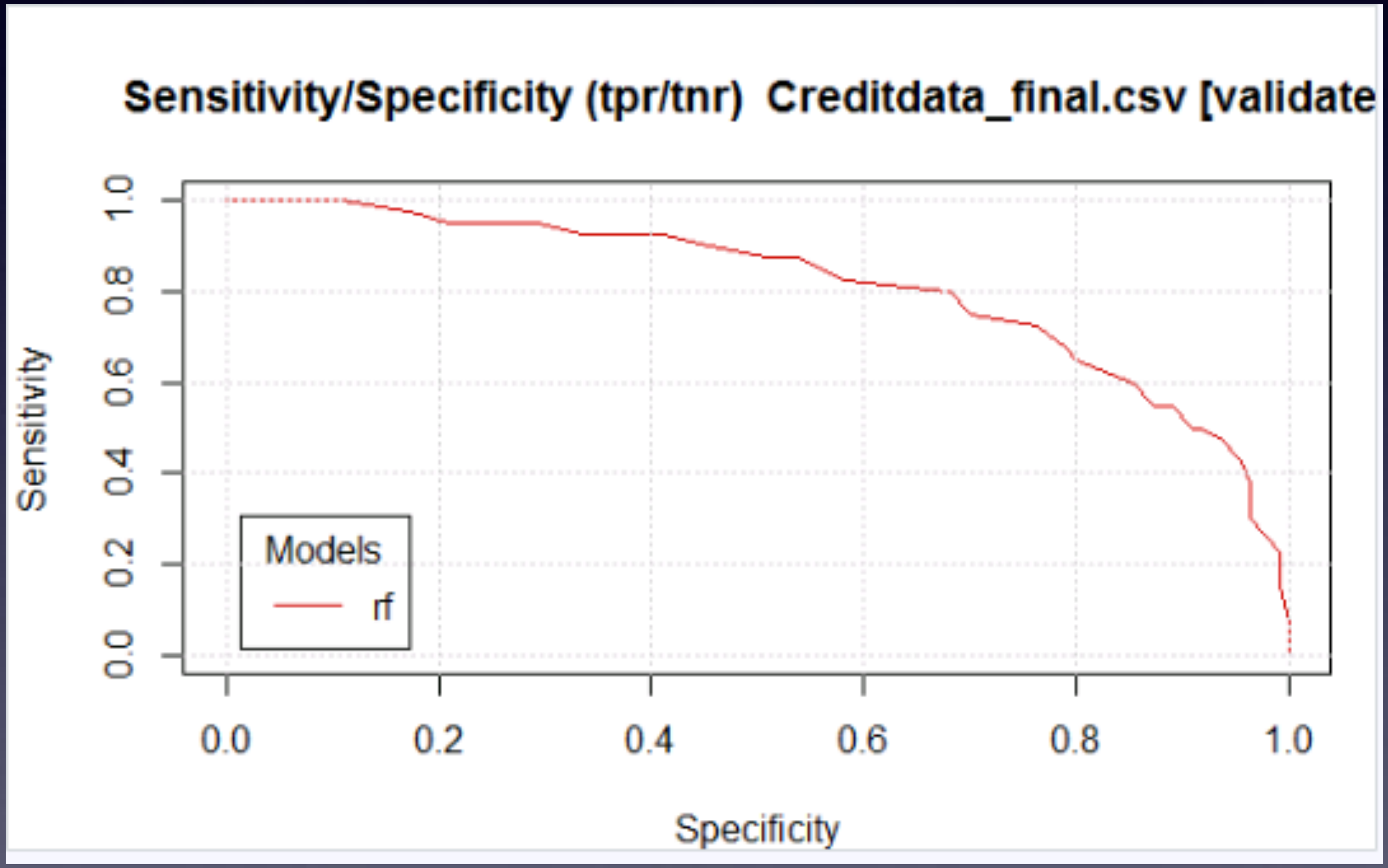
- This error matrix displays the error based upon the test dataset. It displays the disagreement between the final model's predictions and the actual outcomes of the testing observations.

- When referring back to the model summary, we remember the error rate was ~29% for the training set, while this finalized report displays the overall error at 22%, a 7% error decrease. We are now operating at 78% accuracy, as opposed to 70%.

Model Accuracy



Model Accuracy Cont'd



Loss of Value

```
Console  Terminal ×  Jobs ×
~/
> mean(Credit_Data$AMOUNT[Credit_Data$DEFAULT == 1])
[1] 3938.127
```

$3938.13 avg. Amount of Credit

X

300 Applicants

-$590,719.50 in losses if they paid 50% of their obligated credit prior to default

Loss vs ROI

$3938.13 avg. Amount of Credit

X

300 Applicants

-$590,719.50 in losses if they paid 50%

-$590,719.50 with 30% original Default rate

If our model produces 78% accuracy, an 8% increase, in opposition to the original dataset error of default, we then add back 8% in value, or in words, prevent 8% in loss.

Return on Investment

-$590,719.50 in losses if they paid 50% of their original obligated credit, given the 30% original Default rate

+8% improvement in error identification or Default likelihood

$590,719.50 x 0.08 =
$47,257.56 in avoided losses

Day-to-Day Operational Utility

Given the results of the model and the details surrounding variable importance, utilizing such results during day-to-day operations is key

The results of the model in addition to the specifics that evaluate the accuracy of the model, can be utilized by credit branch associates to determine the likelihood that an applicant defaults on their credit

These features can provide substantial increases in operational productivity, with respect to the application processing arena, as well as the efficiency to which risk is more reasonably managed

References

Narkhede, S. (2018). Understanding AUC - ROC curve. Retrieved May 25, 2020 from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

SydneyF. (2018). Alteryx - Help!... Mean Decrease in Gini for dummies. Retrieved May 27, 2020 from https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Help-Mean-Decrease-in-Gini-for-dummies/td-p/197223

Widjaja, J. (2017). How do you explain 'mean decrease accuracy' and 'mean decrease gini' in layman's terms? Retrieved May 27, 2020 from https://www.quora.com/How-do-you-explain-%E2%80%98mean-decrease-accuracy%E2%80%99-and-%E2%80%98mean-decrease-gini%E2%80%99-in-layman%E2%80%99s-terms

Williams, G. (2011). Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discovery. Springer Science+Business Media, LLC. Random Forests. Retrieved May 24, 2020.