Predictive Modeling with DataRobot

by Qi Sun

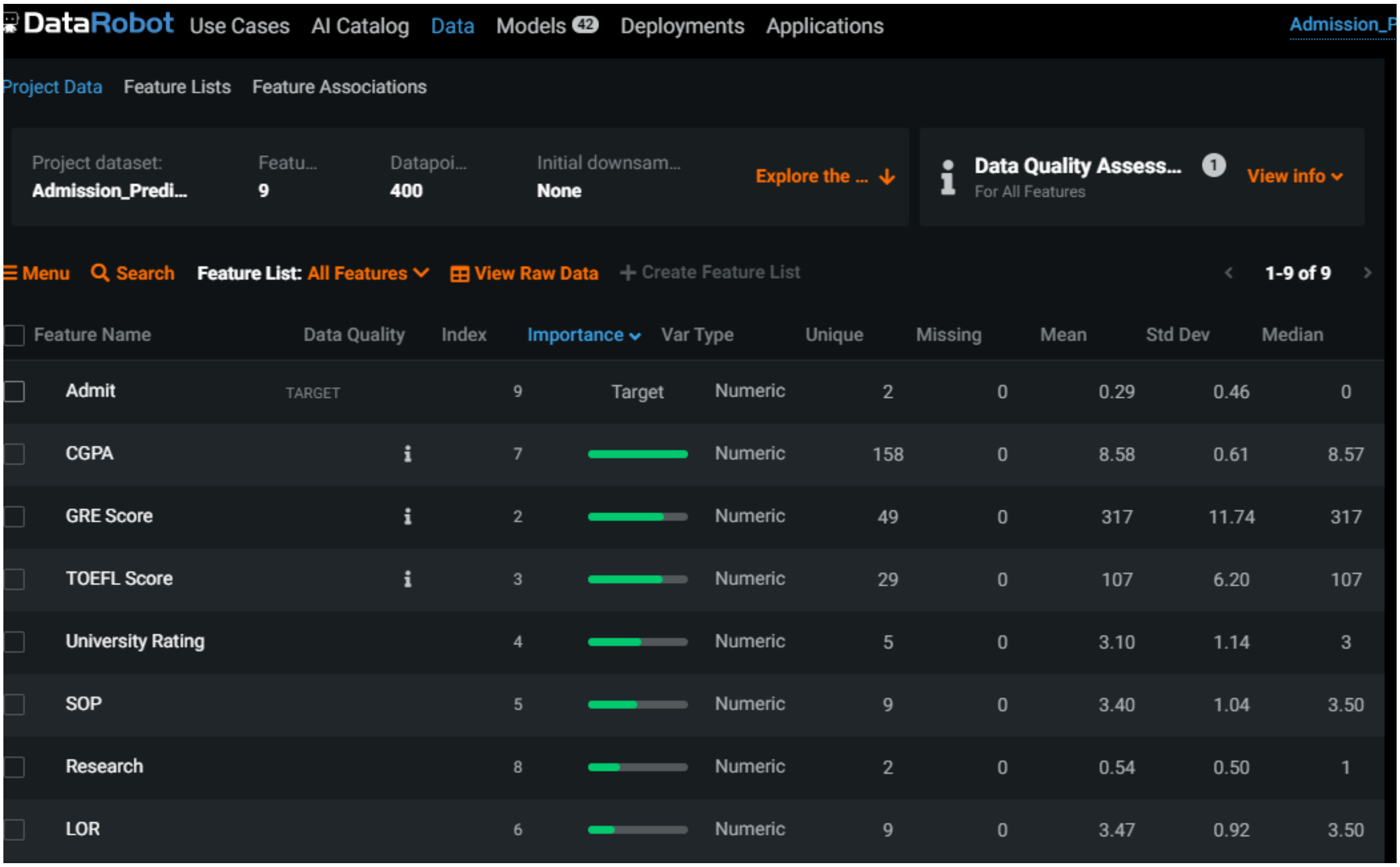
The purpose of this study is to predict Graduate Admissions. The dataset contains several variables which are considered important during the application for Masters Programs.

The independent variables are:

GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose and Letter of Recommendation Strength (out of 5), Undergraduate GPA (out of 10), Research Experience (either 0 or 1).

The dependent variable is Admit (0 or 1). There are a total of 400 records in the dataset.

Here is the description of the dataset:



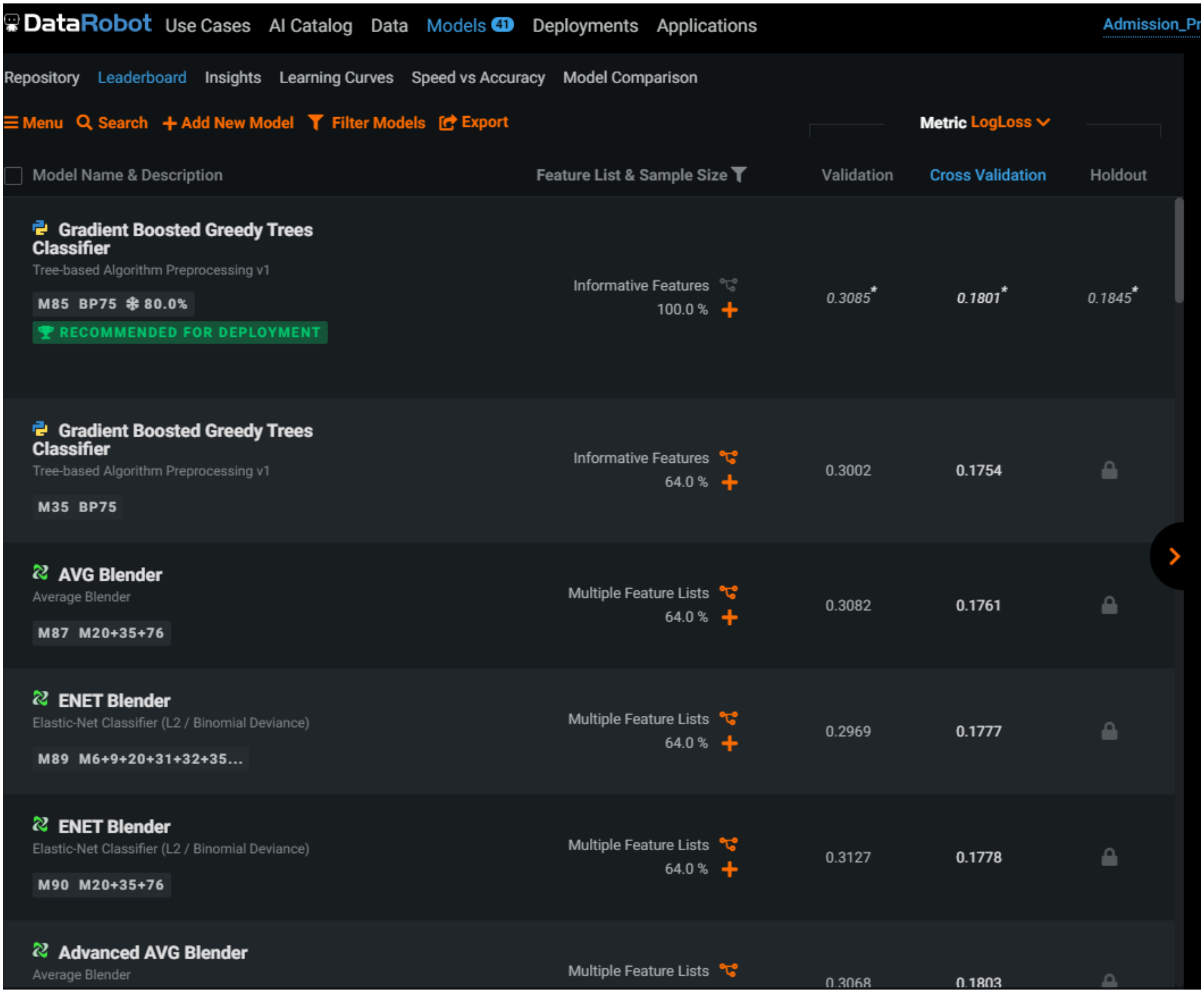
Project dataset: Admission_Predi... Featu... 9 Datapoi... 400 Initial downsam... None

Data Quality Assess... 1 For All Features

Feature List: All Features

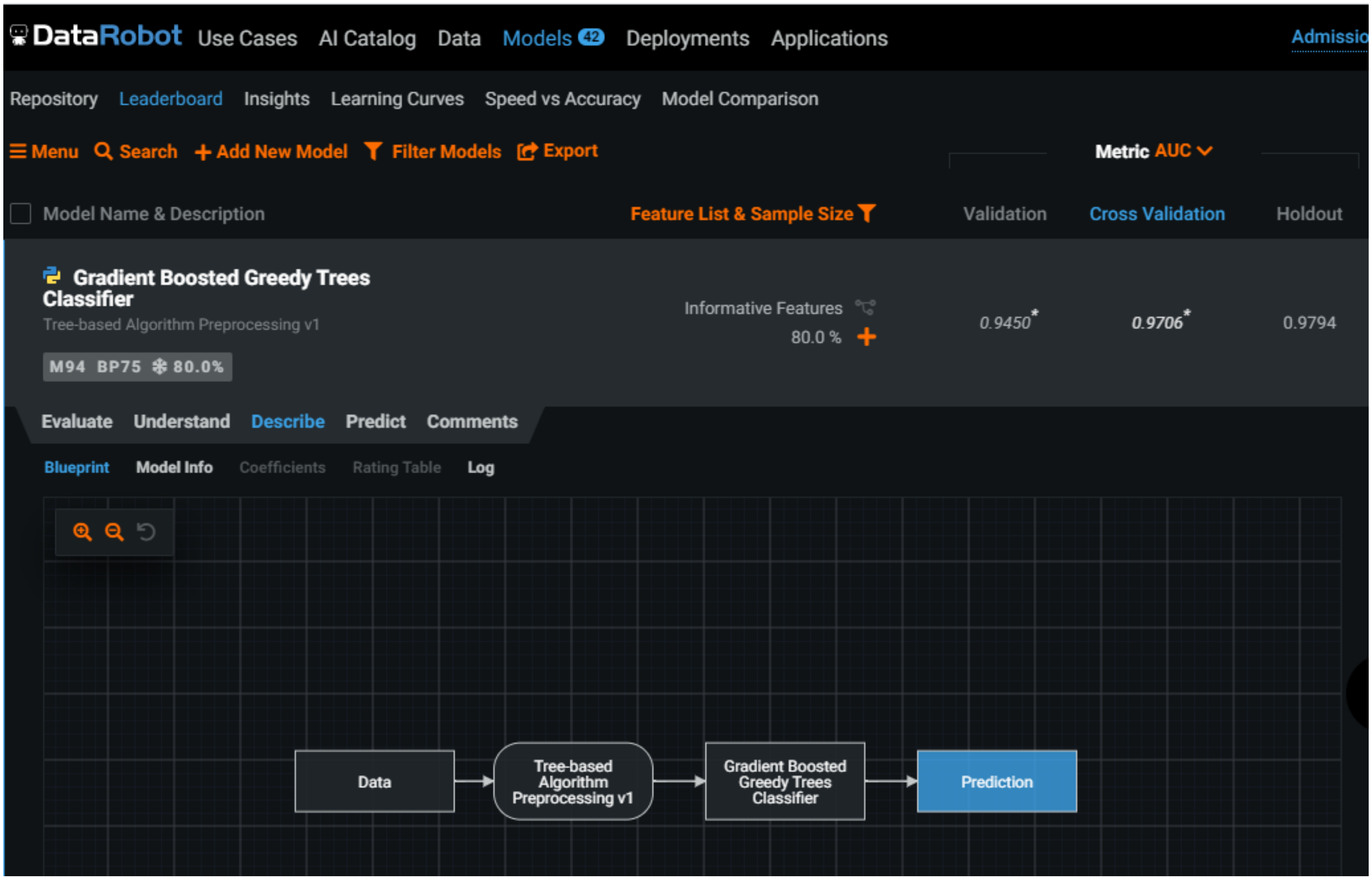
Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median
Admit	TARGET	9	Target	Numeric	2	0	0.29	0.46	0
CGPA	i	7		Numeric	158	0	8.58	0.61	8.57
GRE Score	i	2		Numeric	49	0	317	11.74	317
TOEFL Score	i	3		Numeric	29	0	107	6.20	107
University Rating		4		Numeric	5	0	3.10	1.14	3
SOP		5		Numeric	9	0	3.40	1.04	3.50
Research		8		Numeric	2	0	0.54	0.50	1
LOR		6		Numeric	9	0	3.47	0.92	3.50

The classification prediction was performed by using DataRobot. I selected ‘Autopilot’ on the Modeling Mode. After the process completed, I got the results with LogLoss metric showing below:

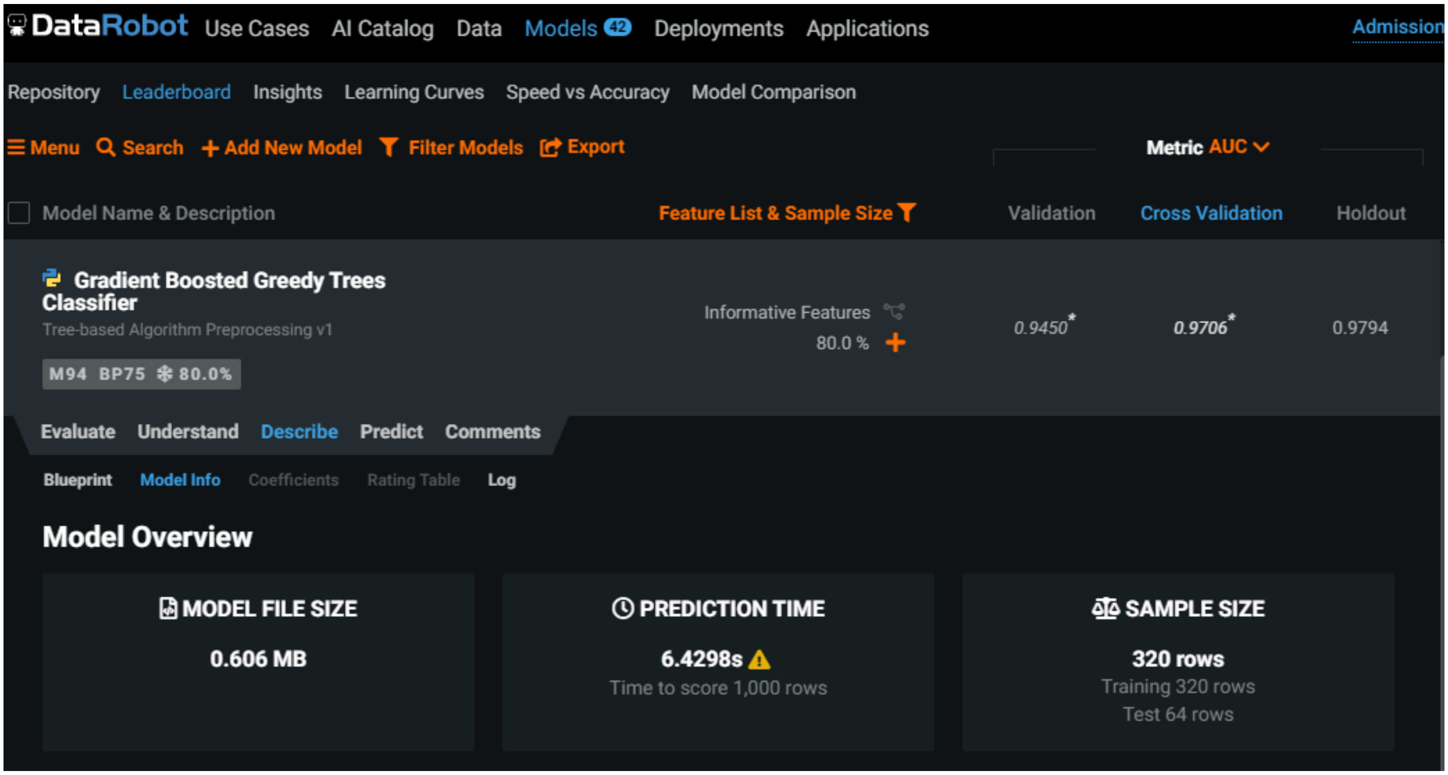


Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Gradient Boosted Greedy Trees Classifier Tree-based Algorithm Preprocessing v1 M85 BP75 ❄ 80.0% RECOMMENDED FOR DEPLOYMENT	Informative Features 100.0 %	0.3085*	0.1801*	0.1845*
Gradient Boosted Greedy Trees Classifier Tree-based Algorithm Preprocessing v1 M35 BP75	Informative Features 64.0 %	0.3002	0.1754	🔒
AVG Blender Average Blender M87 M20+35+76	Multiple Feature Lists 64.0 %	0.3082	0.1761	🔒
ENET Blender Elastic-Net Classifier (L2 / Binomial Deviance) M89 M6+9+20+31+32+35...	Multiple Feature Lists 64.0 %	0.2969	0.1777	🔒
ENET Blender Elastic-Net Classifier (L2 / Binomial Deviance) M90 M20+35+76	Multiple Feature Lists 64.0 %	0.3127	0.1778	🔒
Advanced AVG Blender Average Blender	Multiple Feature Lists	0.3068	0.1803	🔒

From the Model tab, we can see that DataRobot has automatically created 41 models. The recommended model is Gradient Boosted Greedy Trees Classifier with 100% data. I decided to pick this Classifier since it has the highest AUC, F1, and accuracy scores and to improve it with hypertuning. Next, I chose the Gradient Boosted Greedy Trees Classifier with 80% training set. The following screenshot shows a description of the model and the evaluation results by using metric of AUC.



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Gradient Boosted Greedy Trees Classifier Tree-based Algorithm Preprocessing v1 M94 BP75 80.0%	Informative Features 80.0 %	*0.9450**	***0.9706****	0.9794

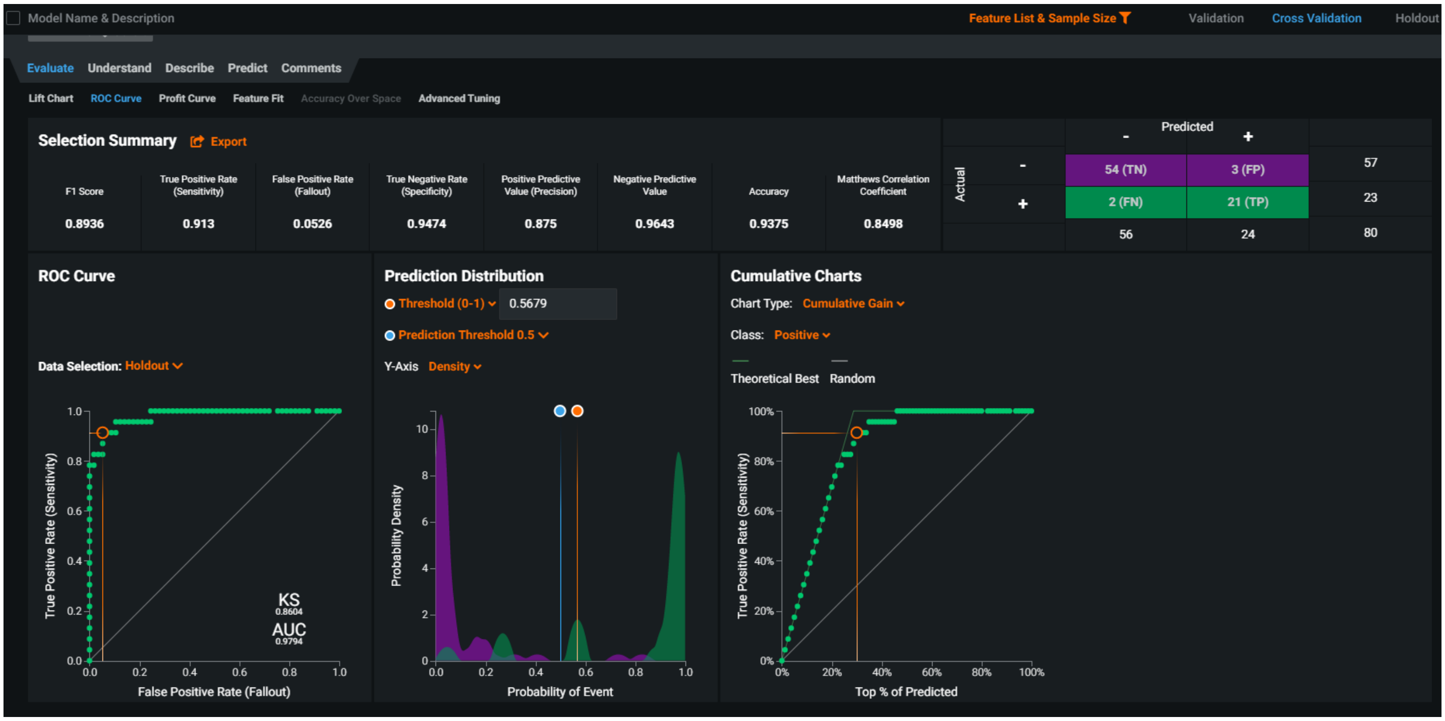


Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Gradient Boosted Greedy Trees Classifier Tree-based Algorithm Preprocessing v1 M94 BP75 80.0%	Informative Features 80.0 %	*0.9450**	***0.9706****	0.9794

Model Overview

MODEL FILE SIZE	PREDICTION TIME	SAMPLE SIZE
0.606 MB	**6.4298s** ⚠ Time to score 1,000 rows	**320 rows** Training 320 rows Test 64 rows

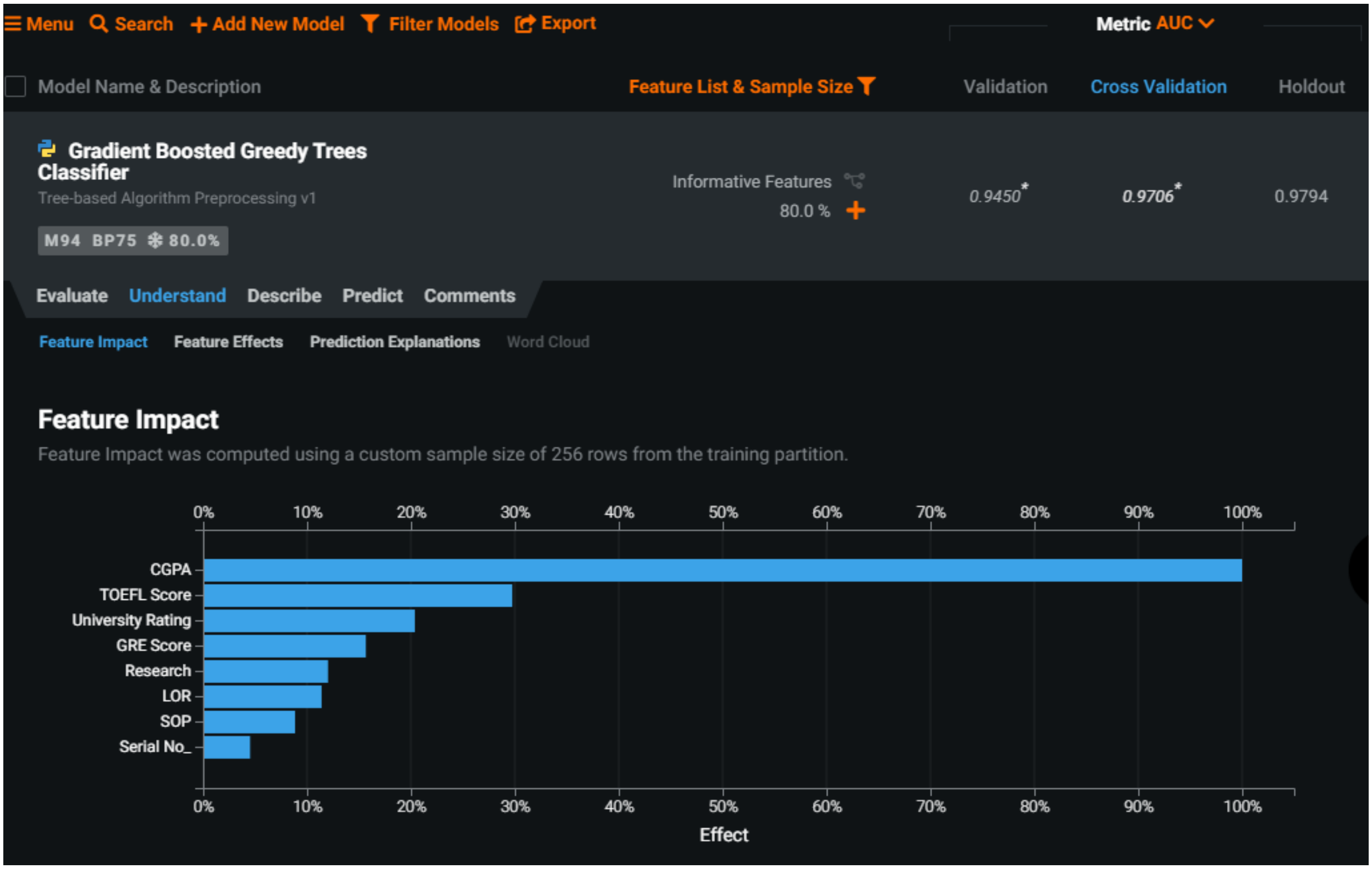
Here is the model evaluation results by using ROC Curve:



F1 Score	True Positive Rate (Sensitivity)	False Positive Rate (Fallout)	True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Negative Predictive Value	Accuracy	Matthews Correlation Coefficient
0.8936	0.913	0.0526	0.9474	0.875	0.9643	0.9375	0.8498

		Predicted -	Predicted +	
Actual	-	54 (TN)	3 (FP)	57
	+	2 (FN)	21 (TP)	23
		56	24	80

Here is the feature impact:



Gradient Boosted Greedy Trees Classifier

Tree-based Algorithm Preprocessing v1

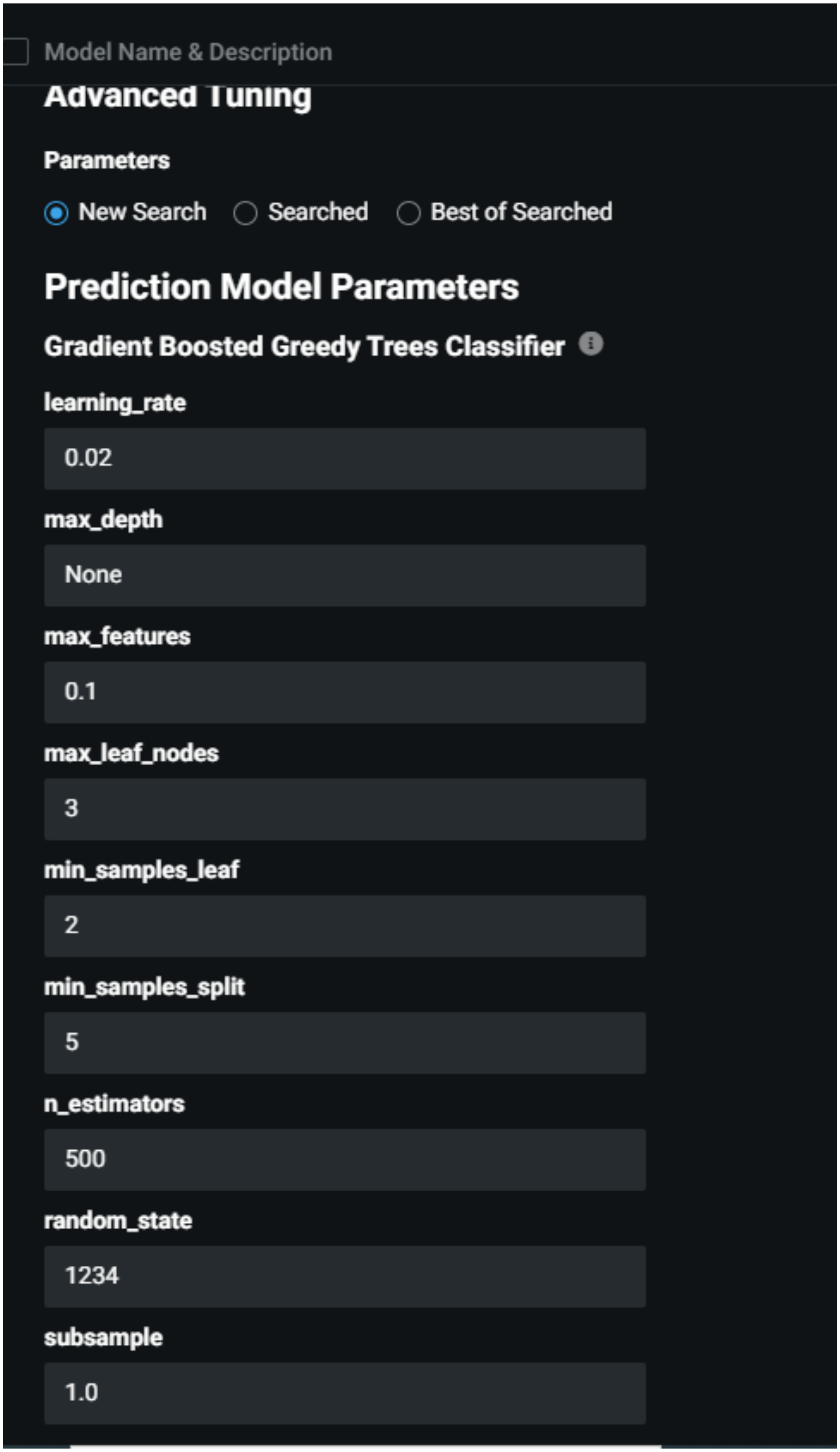
M94 BP75 80.0%

Feature List & Sample Size	Validation	Cross Validation	Holdout
Informative Features 80.0 %	0.9450*	0.9706*	0.9794

Feature Impact

Feature Impact was computed using a custom sample size of 256 rows from the training partition.

Next, I'll do the advanced tuning. The following screenshot shows the values of each parameter on the original Gradient Boosted Greedy Trees Classifier:



□ Model Name & Description

Advanced Tuning

Parameters

● New Search ○ Searched ○ Best of Searched

Prediction Model Parameters

Gradient Boosted Greedy Trees Classifier

learning_rate: 0.02

max_depth: None

max_features: 0.1

max_leaf_nodes: 3

min_samples_leaf: 2

min_samples_split: 5

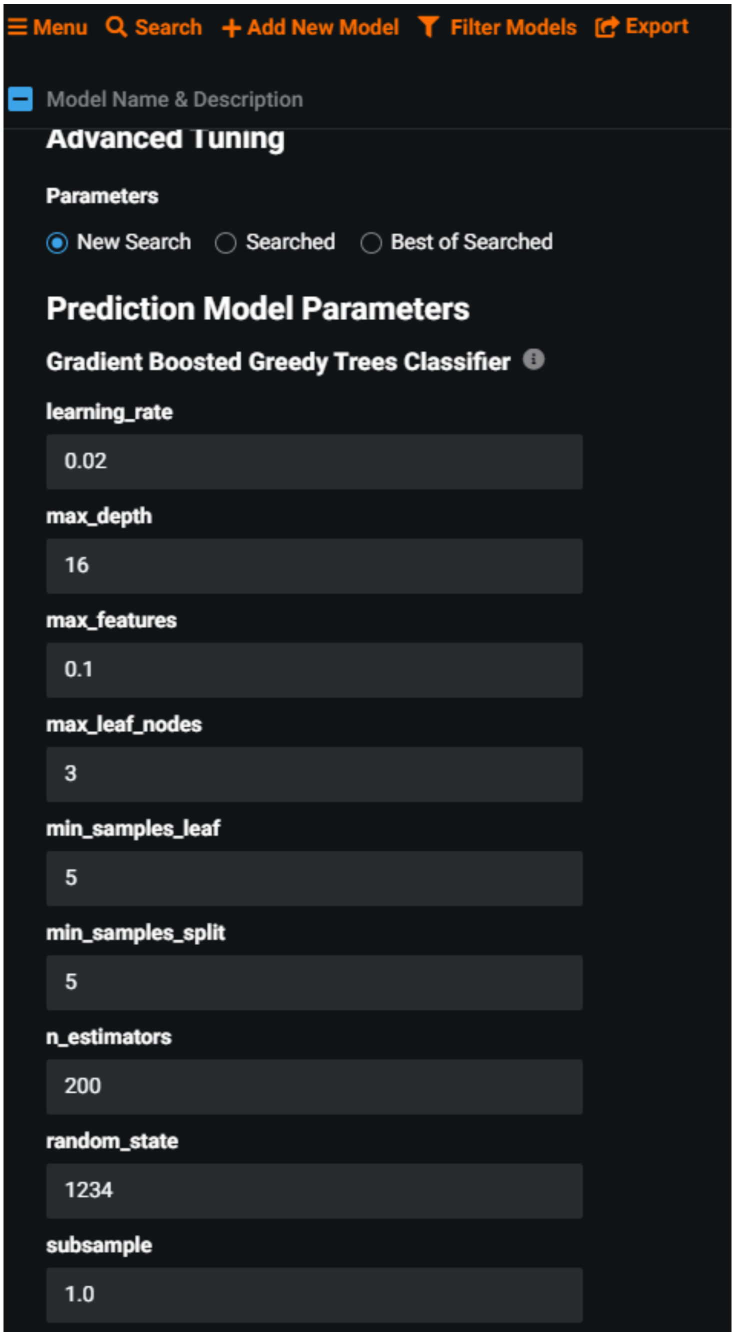
n_estimators: 500

random_state: 1234

subsample: 1.0

Tuning parameters:

I made some changes to max_depth, min_samle_leaf, and n_estimators.



Advanced Tuning

Parameters

● New Search ○ Searched ○ Best of Searched

Prediction Model Parameters

Gradient Boosted Greedy Trees Classifier

learning_rate
0.02

max_depth
16

max_features
0.1

max_leaf_nodes
3

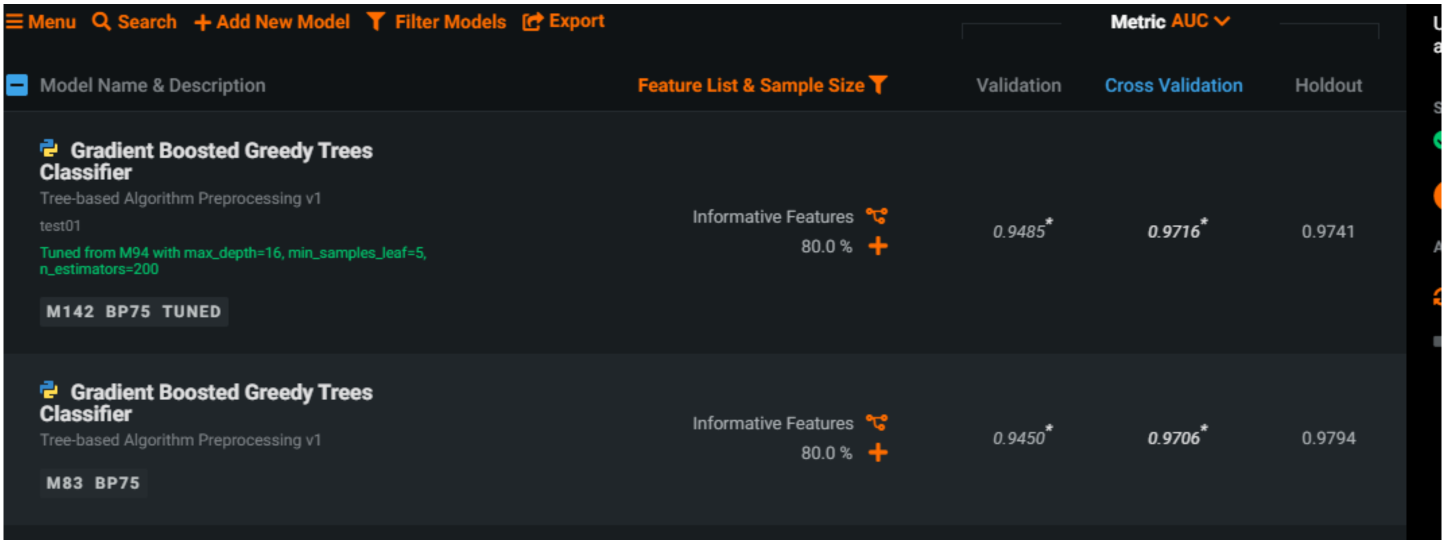
min_samples_leaf
5

min_samples_split
5

n_estimators
200

random_state
1234

subsample
1.0



Metric AUC

Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Gradient Boosted Greedy Trees Classifier Tree-based Algorithm Preprocessing v1 test01 Tuned from M94 with max_depth=16, min_samples_leaf=5, n_estimators=200 M142 BP75 TUNED	Informative Features 80.0 %	0.9485*	0.9716*	0.9741
Gradient Boosted Greedy Trees Classifier Tree-based Algorithm Preprocessing v1 M83 BP75	Informative Features 80.0 %	0.9450*	0.9706*	0.9794

Finally, I got a better AUC scores on the training set (0.9716 vs 0.9706). For this model, the Undergraduate GPA is the most important feature. Next are TOEFL score and GRE score. The Statement of Purpose and Letter of Recommendation Strength is the least important feature in this model.