Machine learning classifier for URLs

Nicolas Ricardo Enciso

UNSECURE

Data set

- 25065 attacks (anomalous) marked as 1
- 36000 normal (normal) marked as 0
- Total 61065 cases in the dataset.

Splitted data

- 70% for training
- 30% for testing
- Random election of the cases from the data
- Two datasets :
    - Original : 8 features with no changes
    - Normalized: numeric values divided by the sum of all values in the case. Xi / sum(Xj)

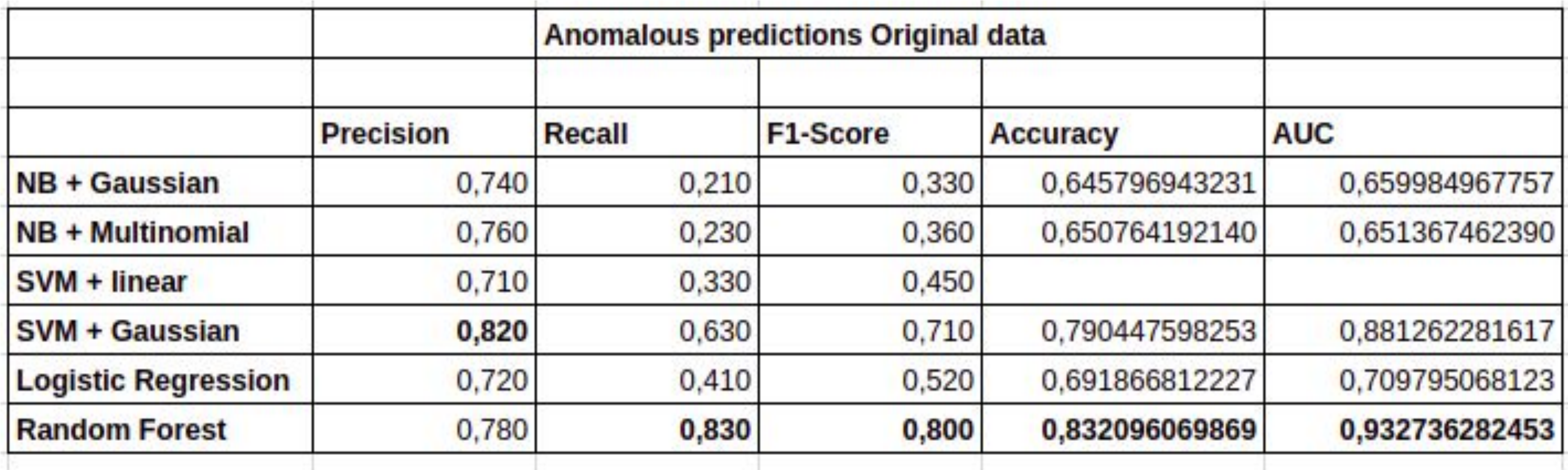
Classifier algorithms

- Naive Bayes classifier:
  - Gaussian classifier
  - Multinomial classifier
- Support Vector Machine:
  - Linear
  - Gaussian (C = 1.11 and gamma = 0.09)
- Logistic Regression
- Random Forest (estimators = 100)

UNSECURE

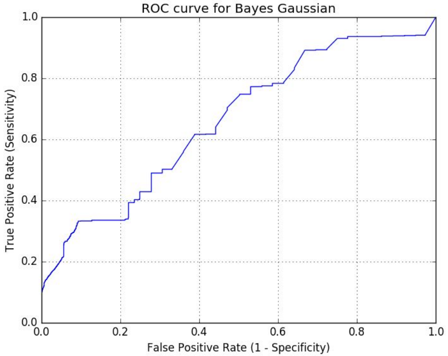
Results original data

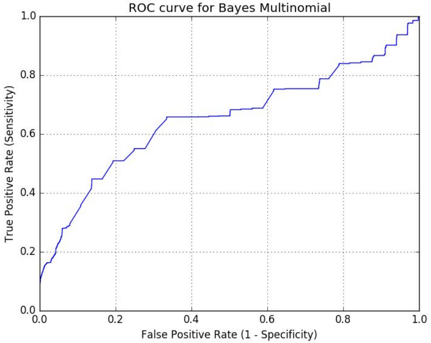
Results performance classification

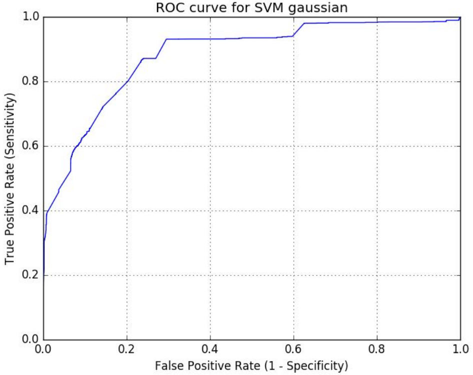


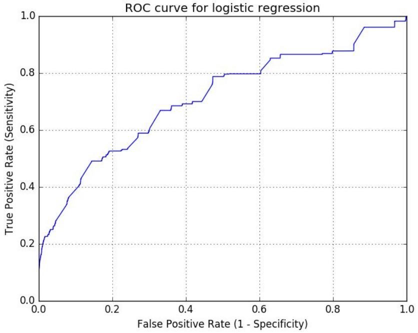
		Anomalous predictions Original data			
	Precision	Recall	F1-Score	Accuracy	AUC
NB + Gaussian	0,740	0,210	0,330	0,645796943231	0,659984967757
NB + Multinomial	0,760	0,230	0,360	0,650764192140	0,651367462390
SVM + linear	0,710	0,330	0,450		
SVM + Gaussian	**0,820**	0,630	0,710	0,790447598253	0,881262281617
Logistic Regression	0,720	0,410	0,520	0,691866812227	0,709795068123
Random Forest	0,780	**0,830**	**0,800**	**0,832096069869**	**0,932736282453**

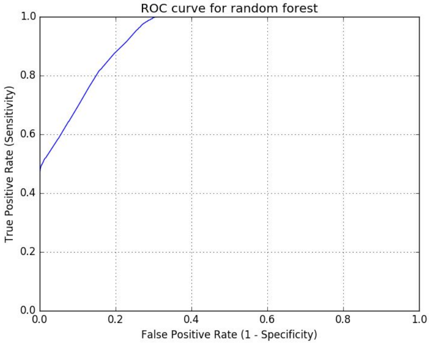
ROC


ROC curve for Bayes Gaussian

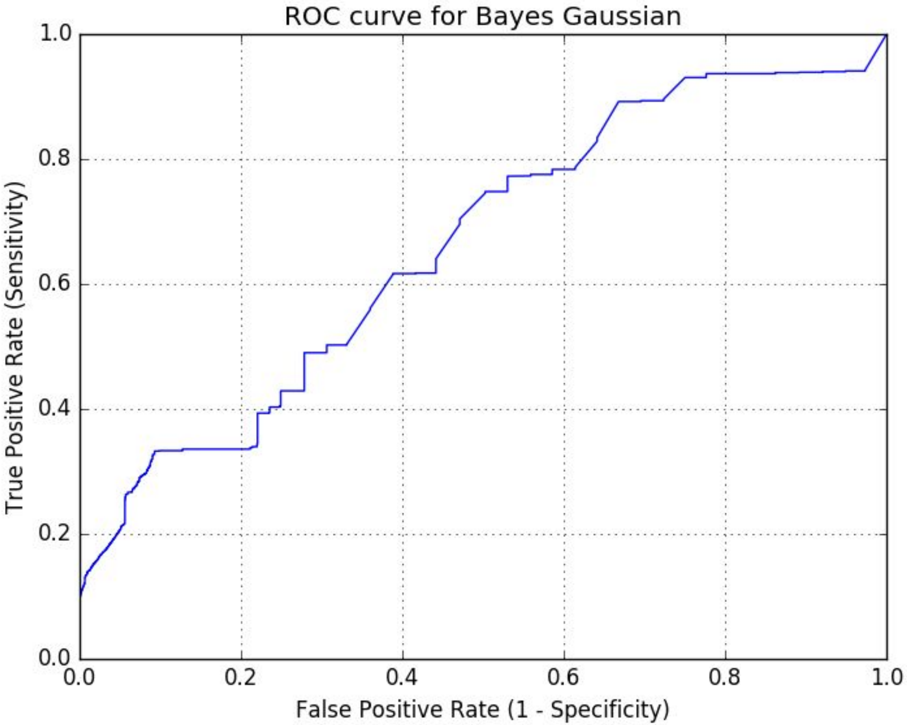

ROC curve for Bayes Multinomial


ROC curve for SVM gaussian

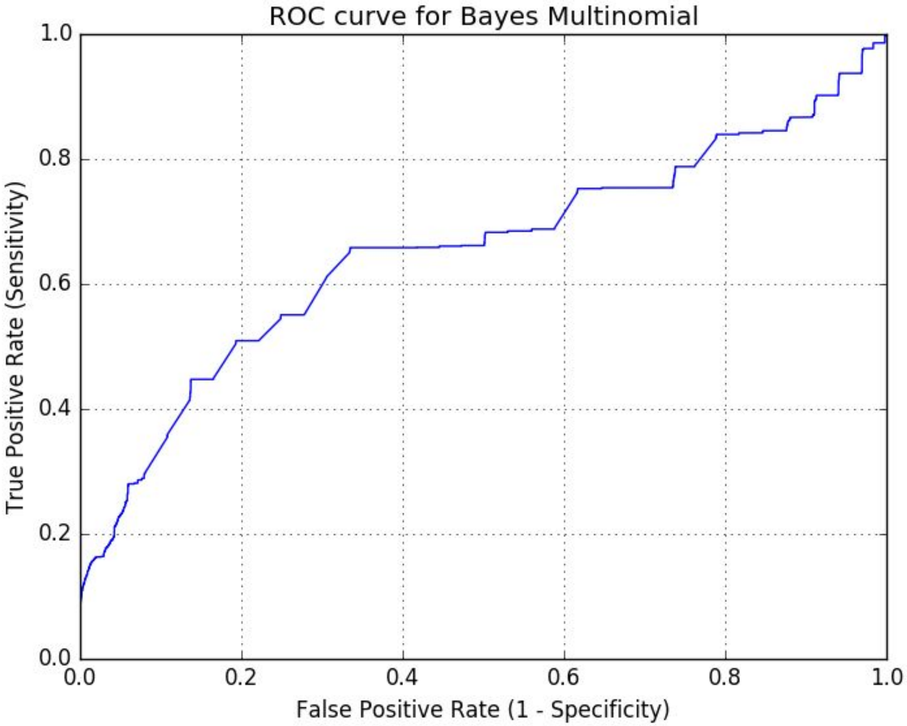

ROC curve for logistic regression


ROC curve for random forest

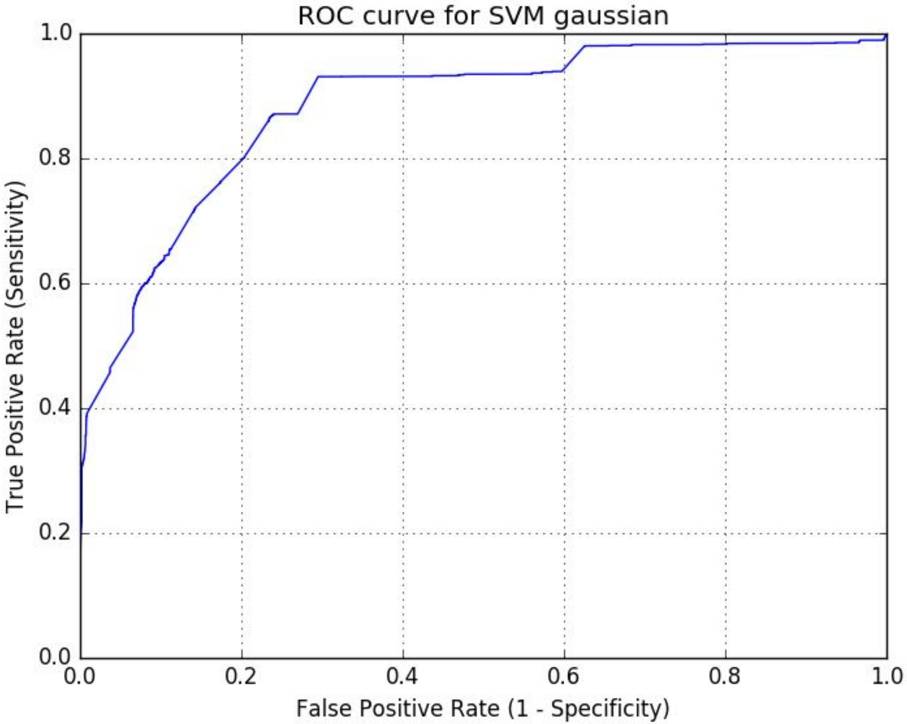
Naive Bayes Gaussian



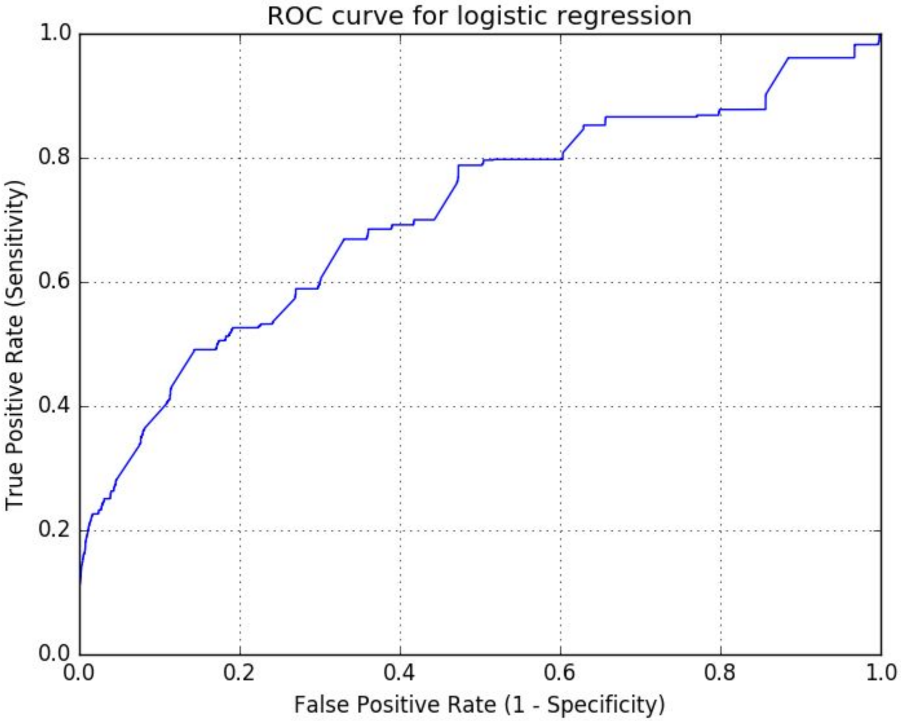
Naive Bayes Multinomial



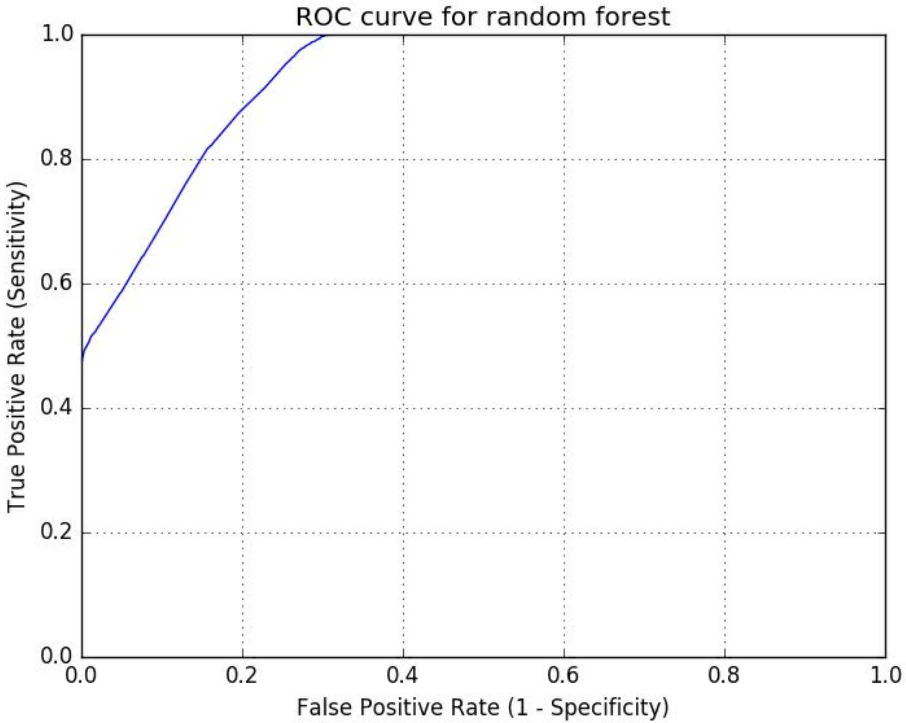
SVM Gaussian



Logistic Regression

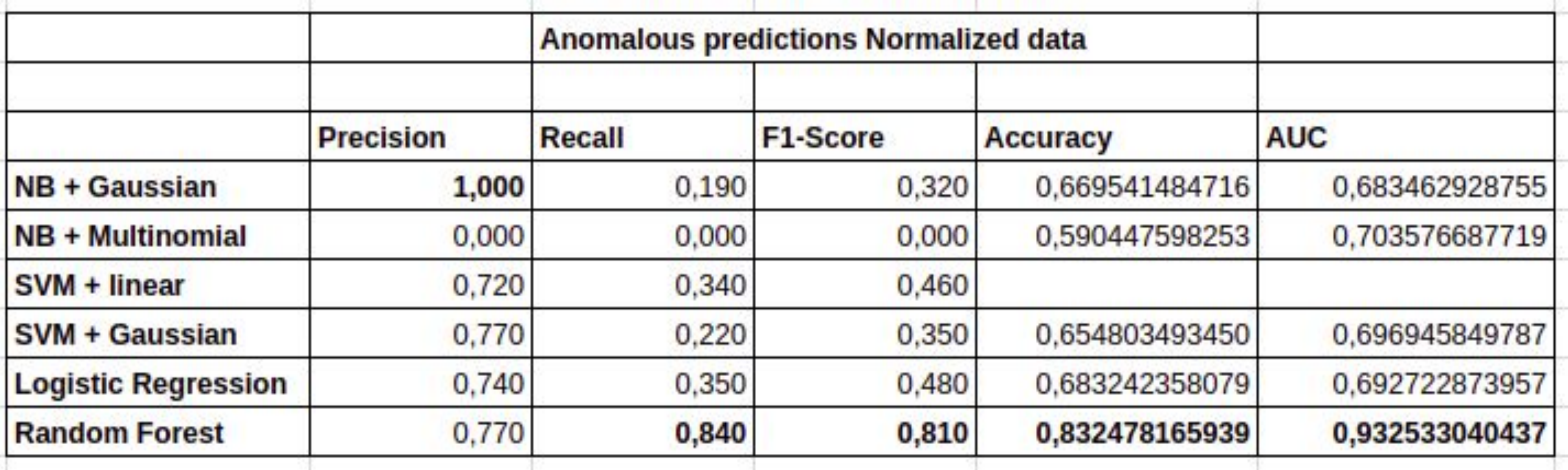


Random forest



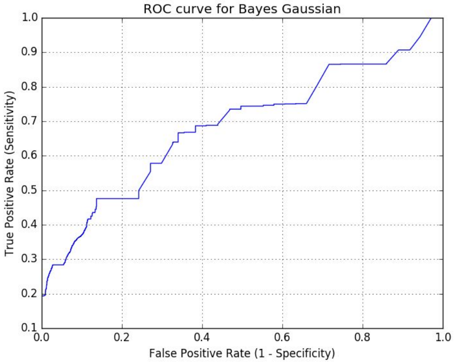
Results normalized data

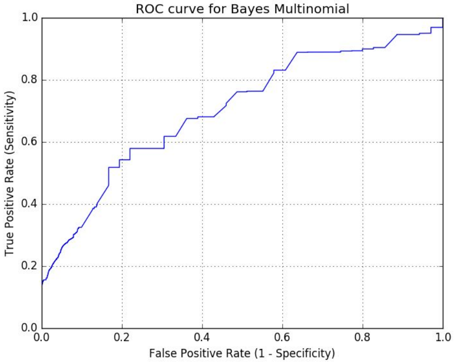
Results performance classification

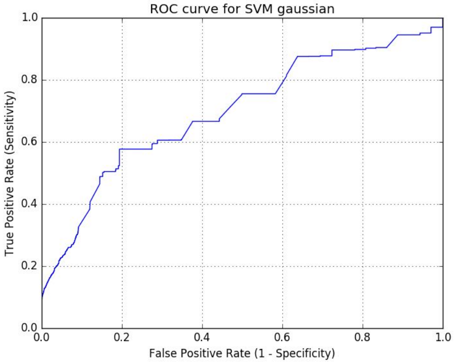


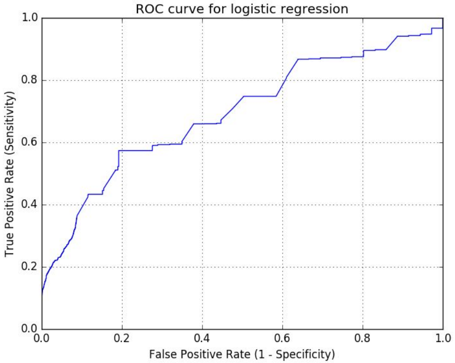
		Anomalous predictions Normalized data			
	Precision	Recall	F1-Score	Accuracy	AUC
NB + Gaussian	**1,000**	0,190	0,320	0,669541484716	0,683462928755
NB + Multinomial	0,000	0,000	0,000	0,590447598253	0,703576687719
SVM + linear	0,720	0,340	0,460		
SVM + Gaussian	0,770	0,220	0,350	0,654803493450	0,696945849787
Logistic Regression	0,740	0,350	0,480	0,683242358079	0,692722873957
Random Forest	0,770	**0,840**	**0,810**	**0,832478165939**	**0,932533040437**

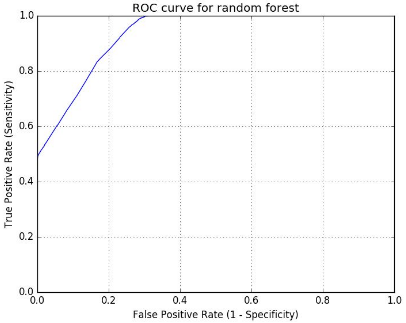
ROC


ROC curve for Bayes Gaussian

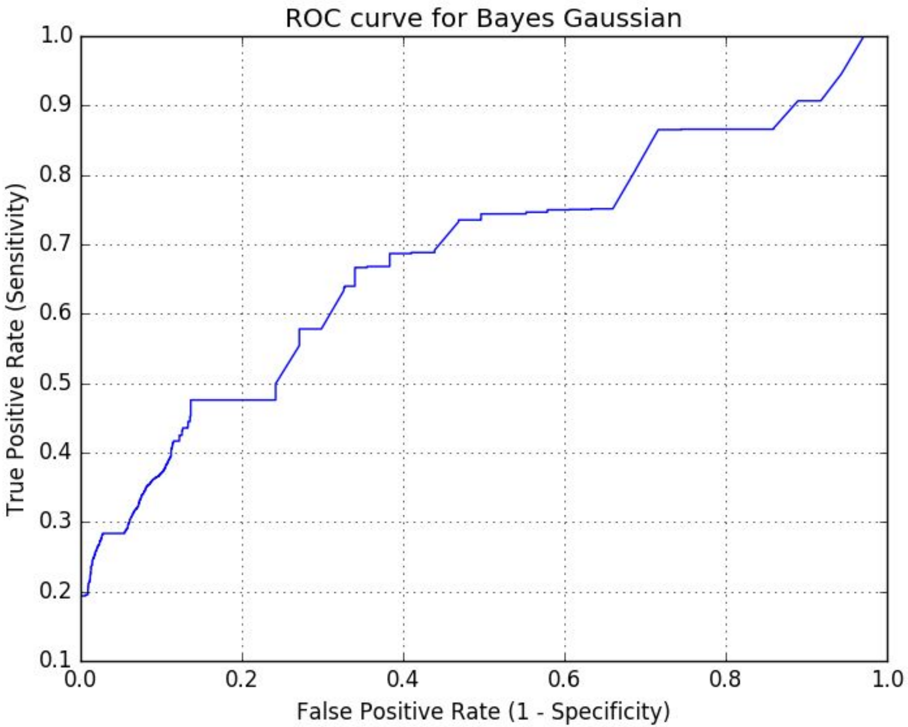

ROC curve for Bayes Multinomial


ROC curve for SVM gaussian

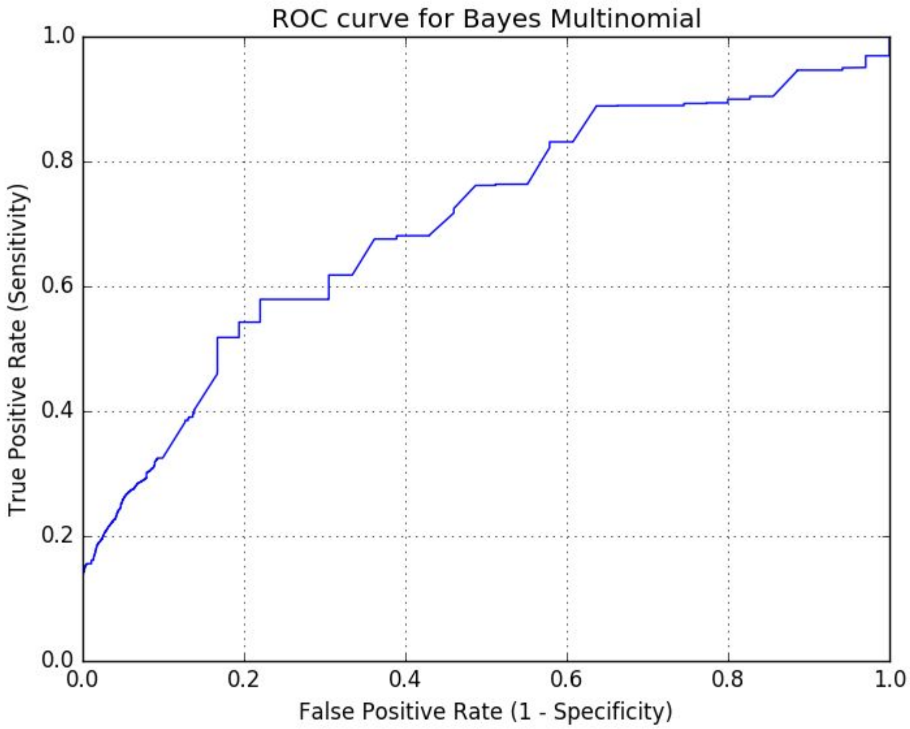

ROC curve for logistic regression


ROC curve for random forest

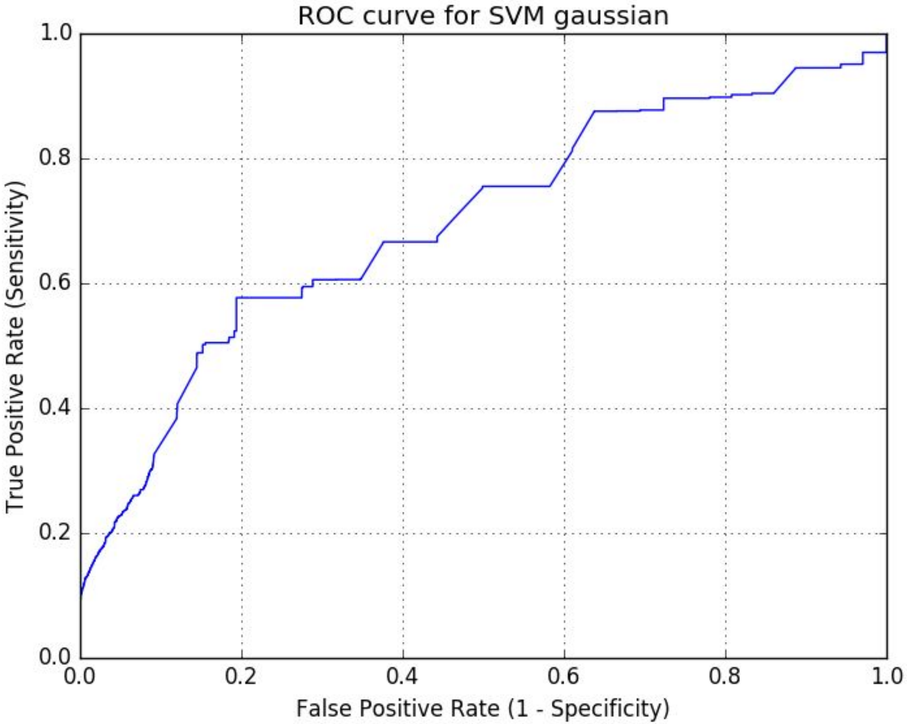
Naive Bayes Gaussian



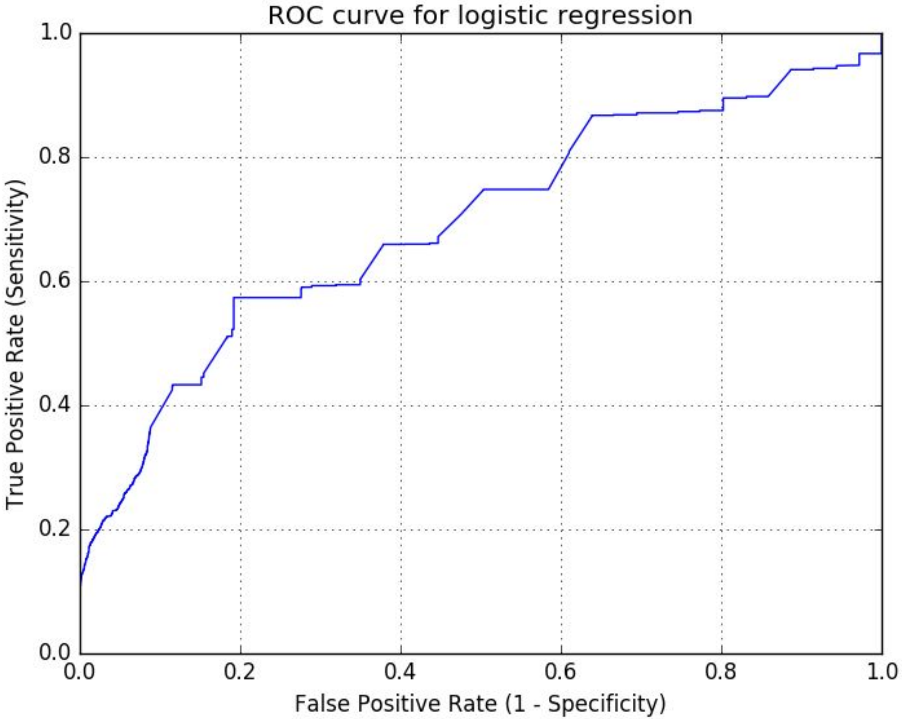
Naive Bayes Multinomial



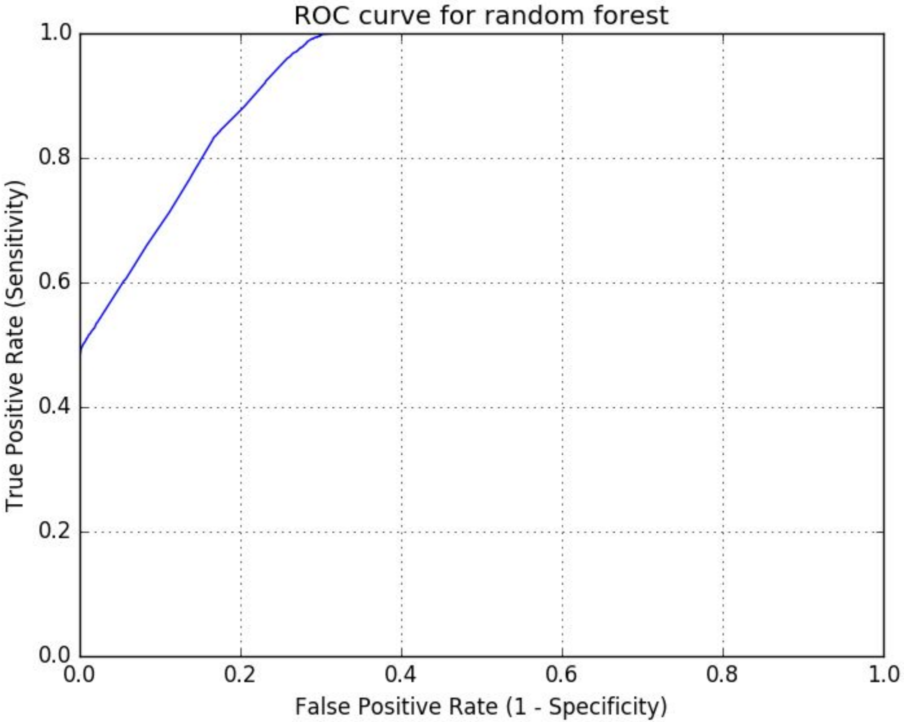
SVM Gaussian



Logistic Regression



Random forest


ROC curve for random forest
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)