Data Glacier

Your Deep Learning Partner

Final Presentation

Project name: Bank Marketing Campaign

Team: Data Science Enthusiasts

Date: August 18th, 2021

Agenda

Executive Summary

Data Understanding

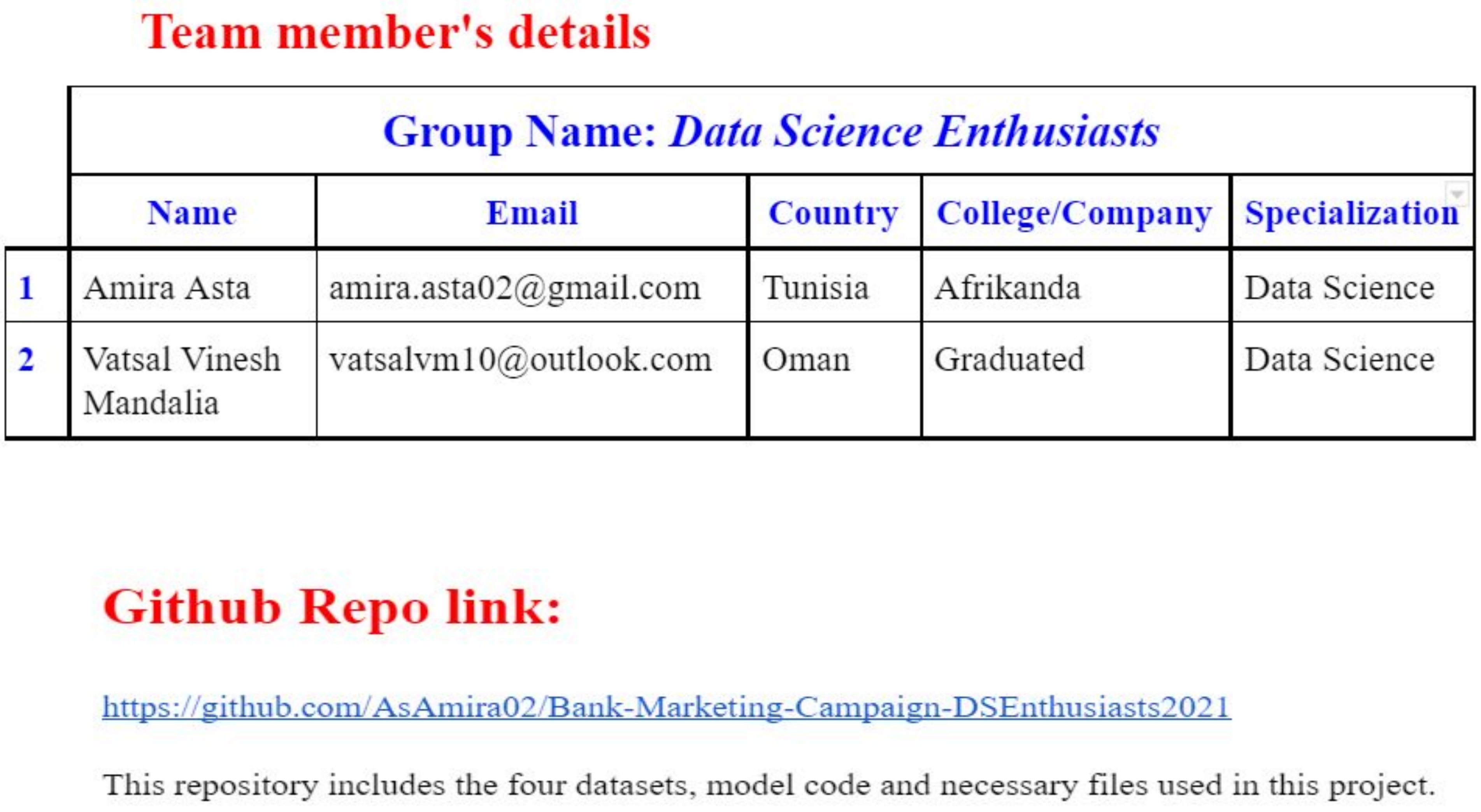
Data Transformation

Data Dependency

Model Building

Model results

Recommendations



Team member's details

Group Name: *Data Science Enthusiasts*					
	Name	**Email**	**Country**	**College/Company**	**Specialization**
1	Amira Asta	amira.asta02@gmail.com	Tunisia	Afrikanda	Data Science
2	Vatsal Vinesh Mandalia	vatsalvm10@outlook.com	Oman	Graduated	Data Science

Github Repo link:

https://github.com/AsAmira02/Bank-Marketing-Campaign-DSEnthusiasts2021

This repository includes the four datasets, model code and necessary files used in this project.

Executive Summary

- **The Client:**

  ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps them understand whether a particular customer will buy their product or not.

- **Problem statement:**

  Build a Classification ML model to shortlist customers who are most likely to buy the term deposit product. This would allow the marketing team to target those customers through various channels.

- **Analysis:**

  The Analysis of this data is divided into the following parts:

  - Data Understanding
  - Univariate analysis
  - Bivariate analysis
  - Model recommendations

Data Understanding

- **<u>Datasets description:</u>**

  Four datasets provided:

  - bank-additional-full: 20 inputs (+1 target variable) and 41119 observations
  - bank-additional: 20 inputs (+1 target variable) and 4119 observations
  - bank-full: 17 inputs (+1 target variable) and 45211 observations
  - bank: 17 inputs (+1 target variable) and 4521 observations

- ➔ The first 7 features are Bank client data.
- ➔ Features 8, 9, 10 and 11 are related to the last contact of the current campaign.
- ➔ Features 12, 13, 14 and 15 are other attributes.
- ➔ Features 16, 17, 18, 19 and 20 are social and economic context attributes.
- ➔ The last feature is the output feature (desired target).

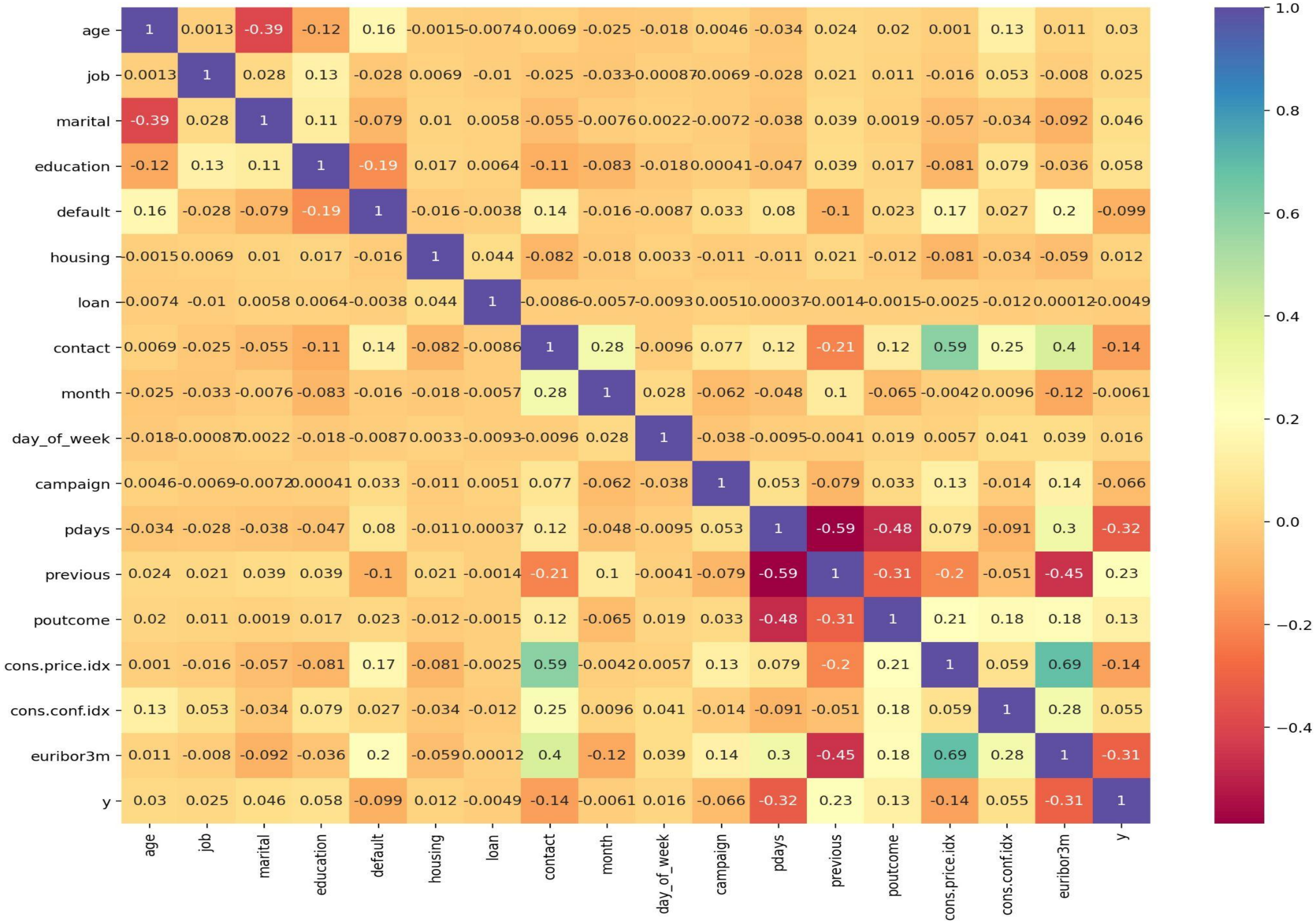
Data Transformation

- **Transformations:**

  - Timeline of observations - May 2008 to November 2010.
  - ‘Duration’ feature is dropped to give realistic predictions from the classification model since it highly effects the output.
  - A frequently occurring missing value ‘unknown’ is considered as another category for the categorical features.
  - Duplicated rows were deleted from the dataset.
  - Outliers were not removed since they will help with the model generalization.
  - Handle highly correlated variables: Observations showed positive high correlation between '**emp.var.rate**', '**nr.employed'**, and '**euribor3m'**. So we drop '**emp.var.rate**' and '**nr.employed'** as **'euribor'** can give us the price of money in the current market.
  - Encoding Categorical features using LabelEncoder since Machine learning algorithms can only read numerical values. It is therefore essential to do this step.

Data Dependency

Now that our dataset contains all numeric variables, we could check correlation between all the features. As shown in the following figure, there are no features that are highly correlated and inversely correlated.



	age	job	marital	education	default	housing	loan	contact	month	day_of_week	campaign	pdays	previous	poutcome	cons.price.idx	cons.conf.idx	euribor3m	y
age	1	0.0013	-0.39	-0.12	0.16	-0.0015	-0.0074	0.0069	-0.025	-0.018	0.0046	-0.034	0.024	0.02	0.001	0.13	0.011	0.03
job	0.0013	1	0.028	0.13	-0.028	0.0069	-0.01	-0.025	-0.033	-0.00087	-0.0069	-0.028	0.021	0.011	-0.016	0.053	-0.008	0.025
marital	-0.39	0.028	1	0.11	-0.079	0.01	0.0058	-0.055	-0.0076	0.0022	-0.0072	-0.038	0.039	0.0019	-0.057	-0.034	-0.092	0.046
education	-0.12	0.13	0.11	1	-0.19	0.017	0.0064	-0.11	-0.083	-0.018	0.00041	-0.047	0.039	0.017	-0.081	0.079	-0.036	0.058
default	0.16	-0.028	-0.079	-0.19	1	-0.016	-0.0038	0.14	-0.016	-0.0087	0.033	0.08	-0.1	0.023	0.17	0.027	0.2	-0.099
housing	-0.0015	0.0069	0.01	0.017	-0.016	1	0.044	-0.082	-0.018	0.0033	-0.011	-0.011	0.021	-0.012	-0.081	-0.034	-0.059	0.012
loan	-0.0074	-0.01	0.0058	0.0064	-0.0038	0.044	1	-0.0086	-0.0057	-0.0093	0.0051	0.00037	-0.0014	-0.0015	-0.0025	-0.012	0.00012	-0.0049
contact	0.0069	-0.025	-0.055	-0.11	0.14	-0.082	-0.0086	1	0.28	-0.0096	0.077	0.12	-0.21	0.12	0.59	0.25	0.4	-0.14
month	-0.025	-0.033	-0.0076	-0.083	-0.016	-0.018	-0.0057	0.28	1	0.028	-0.062	-0.048	0.1	-0.065	-0.0042	0.0096	-0.12	-0.0061
day_of_week	-0.018	-0.00087	0.0022	-0.018	-0.0087	0.0033	-0.0093	-0.0096	0.028	1	-0.038	-0.0095	-0.0041	0.019	0.0057	0.041	0.039	0.016
campaign	0.0046	-0.0069	-0.0072	0.00041	0.033	-0.011	0.0051	0.077	-0.062	-0.038	1	0.053	-0.079	0.033	0.13	-0.014	0.14	-0.066
pdays	-0.034	-0.028	-0.038	-0.047	0.08	-0.011	0.00037	0.12	-0.048	-0.0095	0.053	1	-0.59	-0.48	0.079	-0.091	0.3	-0.32
previous	0.024	0.021	0.039	0.039	-0.1	0.021	-0.0014	-0.21	0.1	-0.0041	-0.079	-0.59	1	-0.31	-0.2	-0.051	-0.45	0.23
poutcome	0.02	0.011	0.0019	0.017	0.023	-0.012	-0.0015	0.12	-0.065	0.019	0.033	-0.48	-0.31	1	0.21	0.18	0.18	0.13
cons.price.idx	0.001	-0.016	-0.057	-0.081	0.17	-0.081	-0.0025	0.59	-0.0042	0.0057	0.13	0.079	-0.2	0.21	1	0.059	0.69	-0.14
cons.conf.idx	0.13	0.053	-0.034	0.079	0.027	-0.034	-0.012	0.25	0.0096	0.041	-0.014	-0.091	-0.051	0.18	0.059	1	0.28	0.055
euribor3m	0.011	-0.008	-0.092	-0.036	0.2	-0.059	0.00012	0.4	-0.12	0.039	0.14	0.3	-0.45	0.18	0.69	0.28	1	-0.31
y	0.03	0.025	0.046	0.058	-0.099	0.012	-0.0049	-0.14	-0.0061	0.016	-0.066	-0.32	0.23	0.13	-0.14	0.055	-0.31	1

Model Building

In order to predict the client subscription for a deposit term, we will use a predictive ML model to help us identify potential customers. As a start we begin by splitting the dataset, into training and testing sets in **80%** and **20%** respectively.

We choose to test out the following set of models since we don't know yet what algorithms will do well on this dataset.

The following algorithms selected for this classification problem include:

- **Linear Algorithms:**

  Logistic Regression (LR) ***(Base Model)***

  Linear Discriminant Analysis (LDA).

- **Nonlinear Algorithms:**

  Classification and Regression Trees (CART),

  Support Vector Machines (SVM),

  Gaussian Naive Bayes (NB)

  k-Nearest Neighbors (KNN).

- **Ensemble Methods:**

  Boosting Methods: AdaBoost (AB) and Gradient Boosting (GBM).

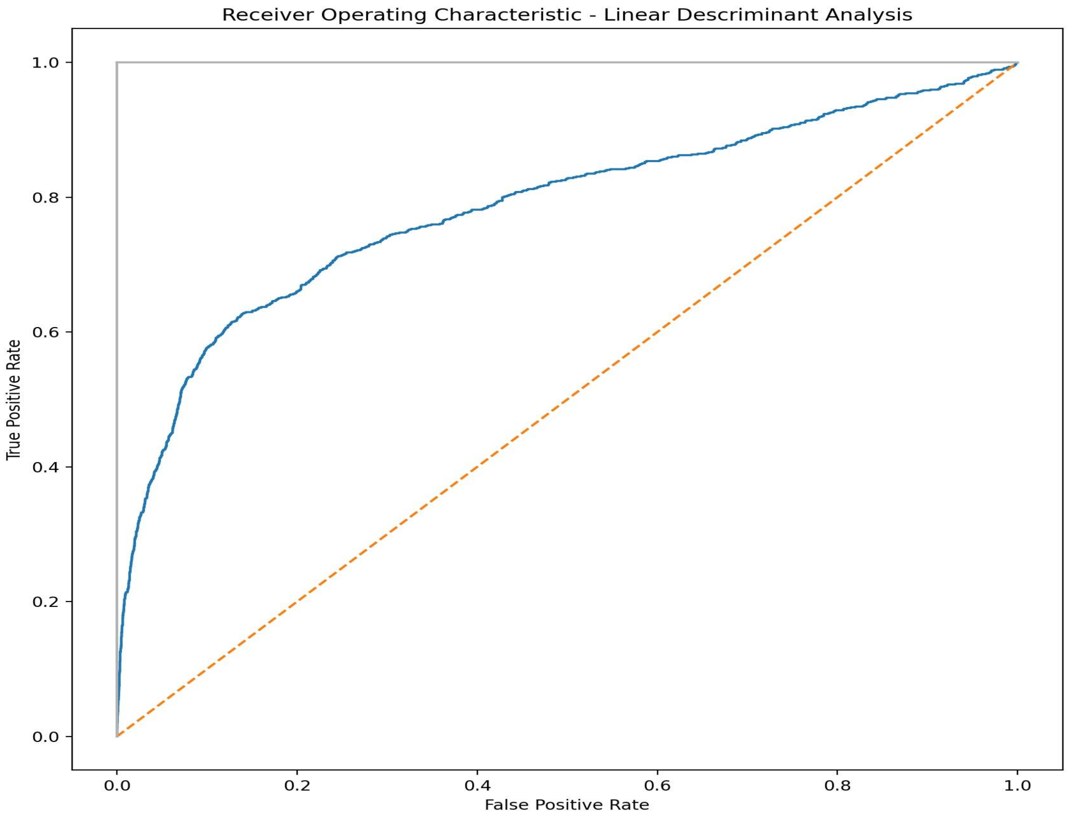
  Bagging Methods: Random Forests (RF) and Extra Trees (ET).

Model Results

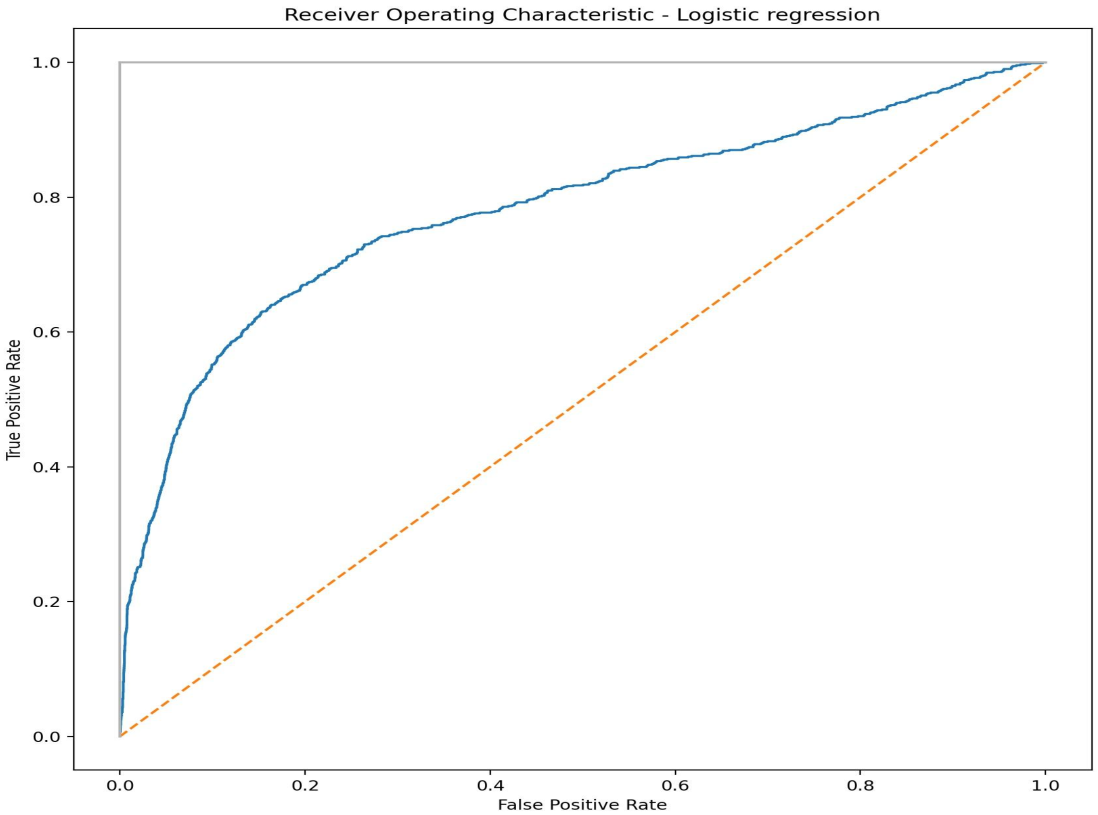
Model results for Linear and nonlinear models

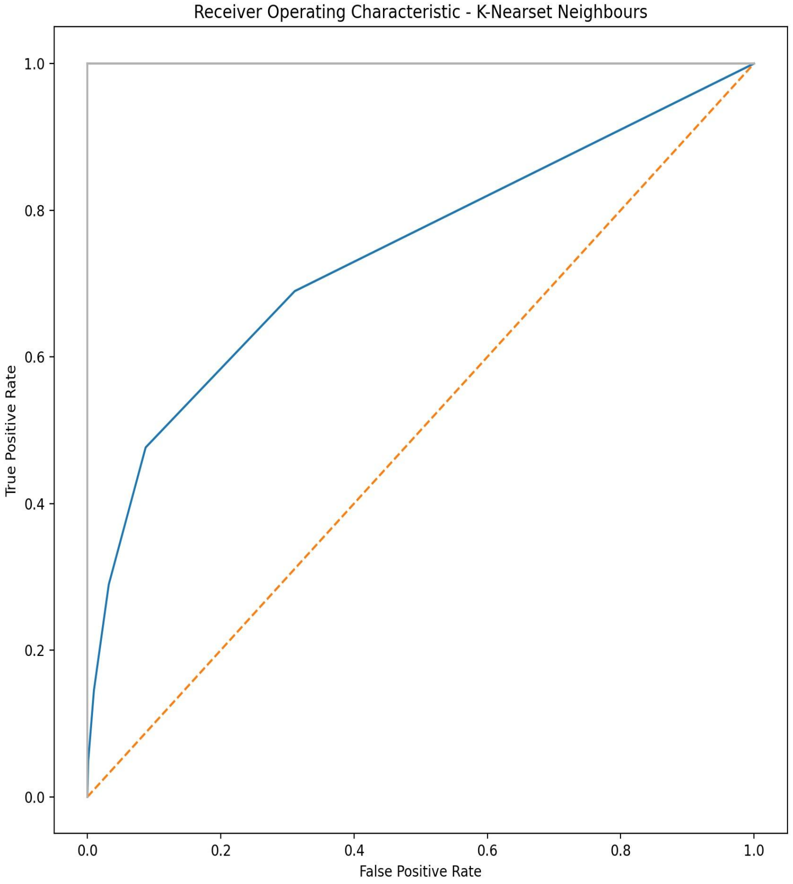
Logistic Regression is used as a base model in our case. As shown in the table, LDA gave better results than LR. Here are the ROC AUC Curves plotted separately for each model.

Linear Discriminant Analysis



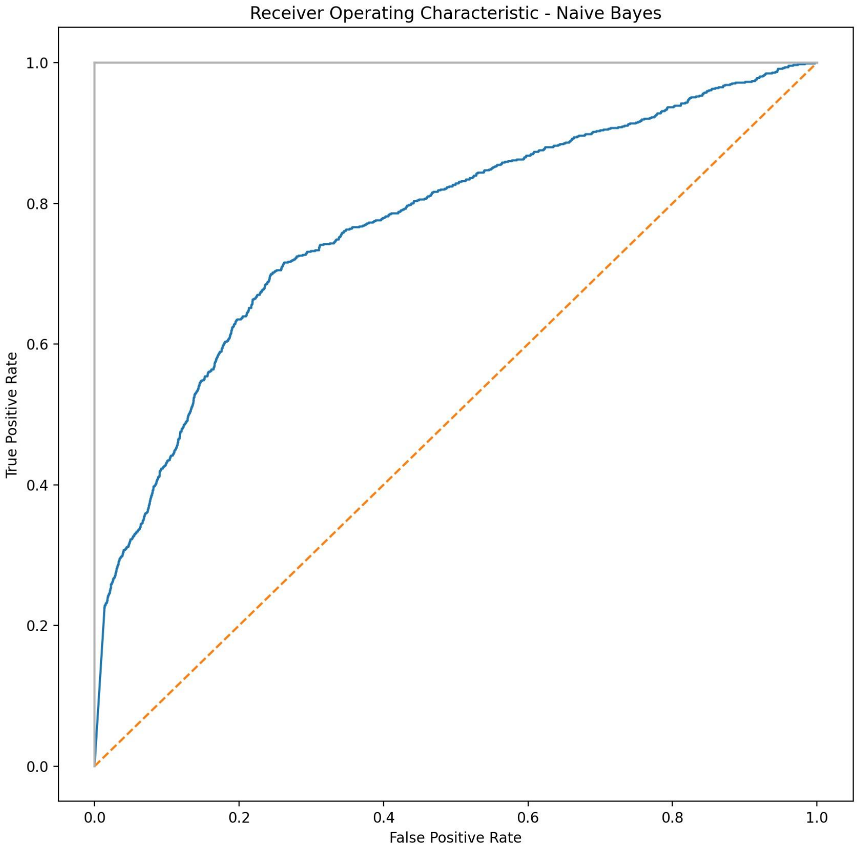
Logistic Regression

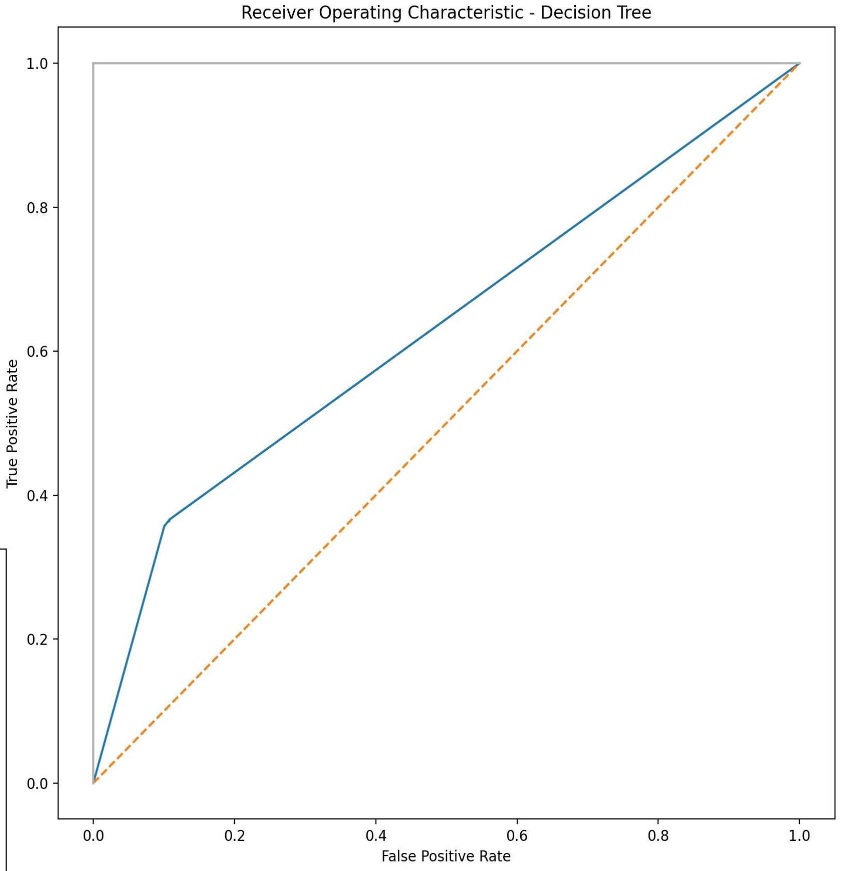




KNN

Naive Bayes

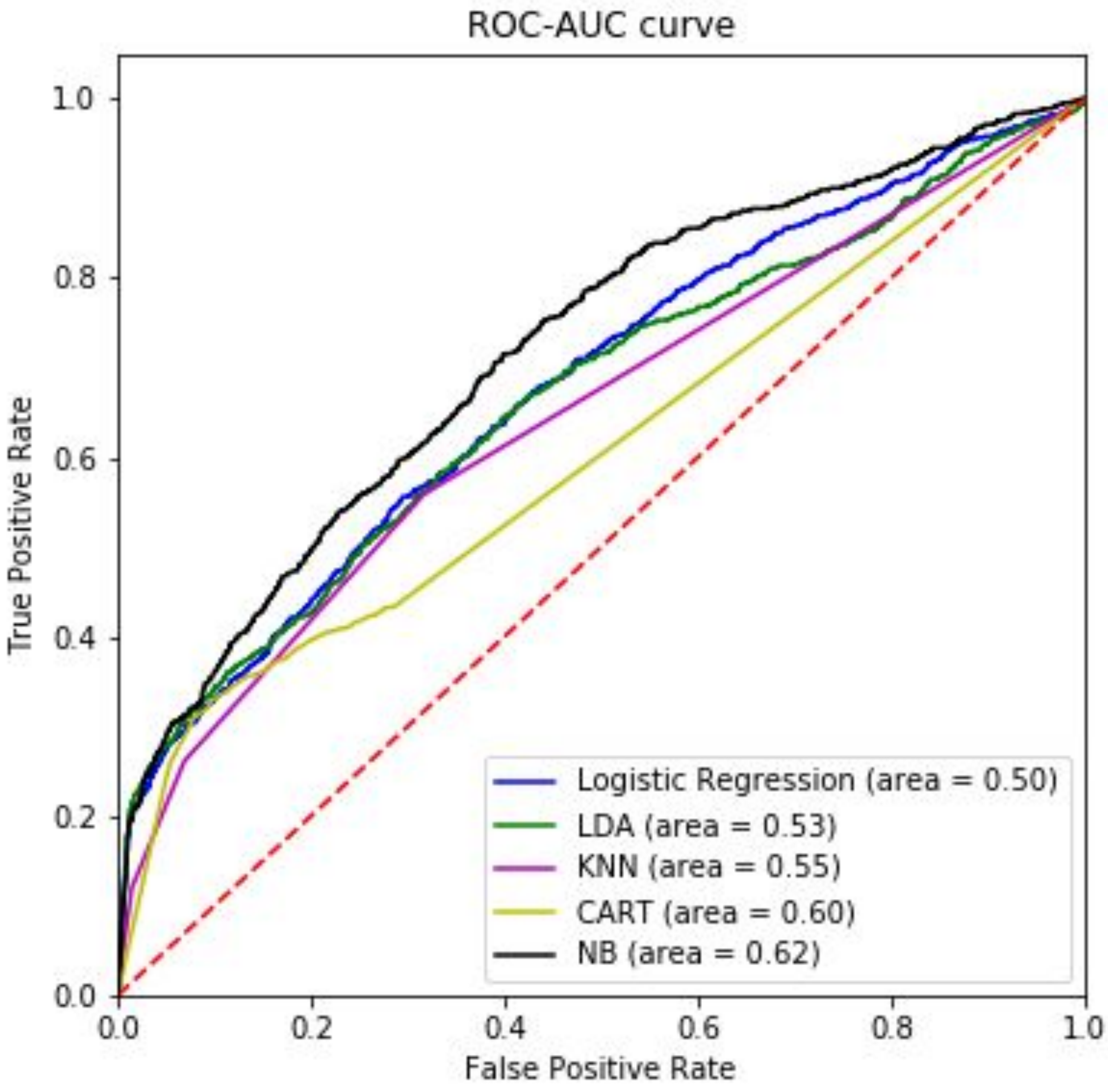




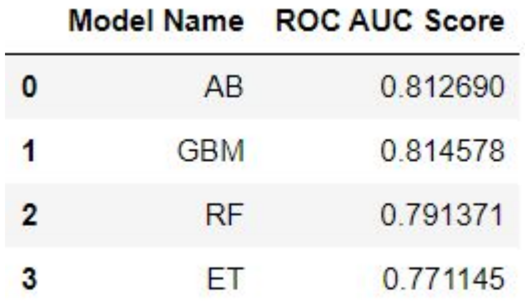
Decision Tree

The above results were gathered in one plot to compare the models.

The ROC curves are produced without Standard Scaler and Cross Validation techniques. This gives a basic result.

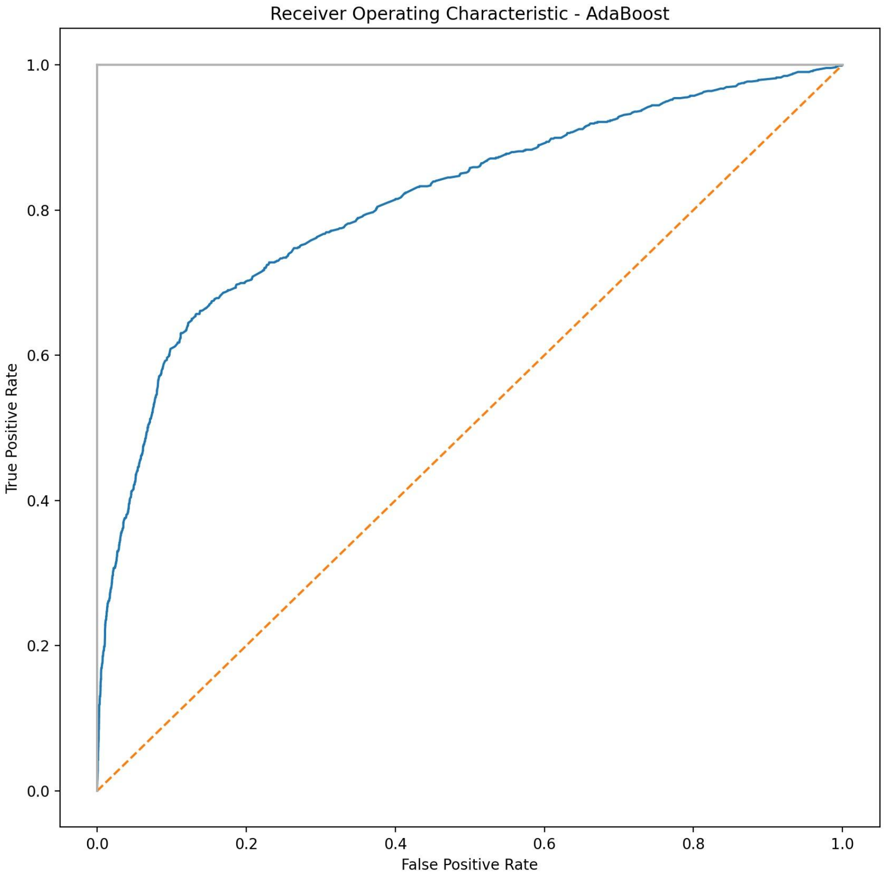


Model results for Ensemble methods

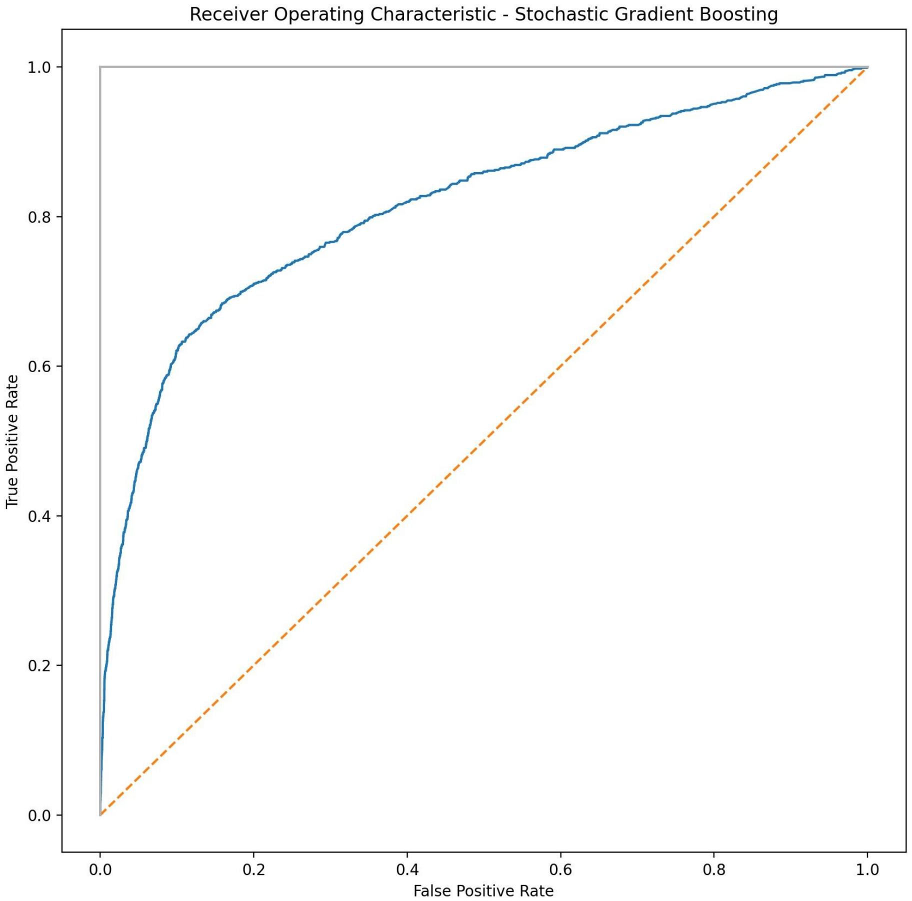


	Model Name	ROC AUC Score
0	AB	0.812690
1	GBM	0.814578
2	RF	0.791371
3	ET	0.771145

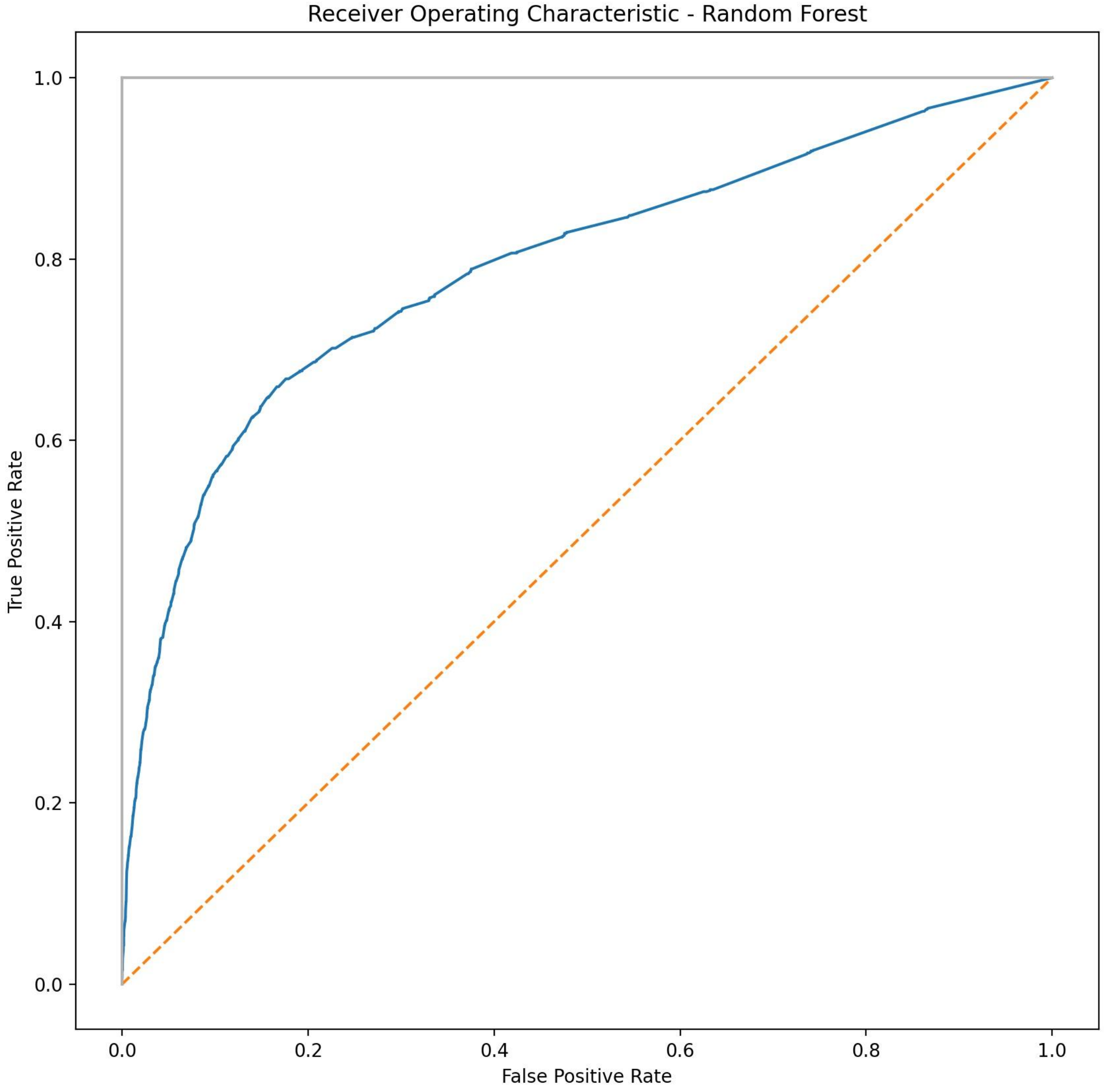
AdaBoost

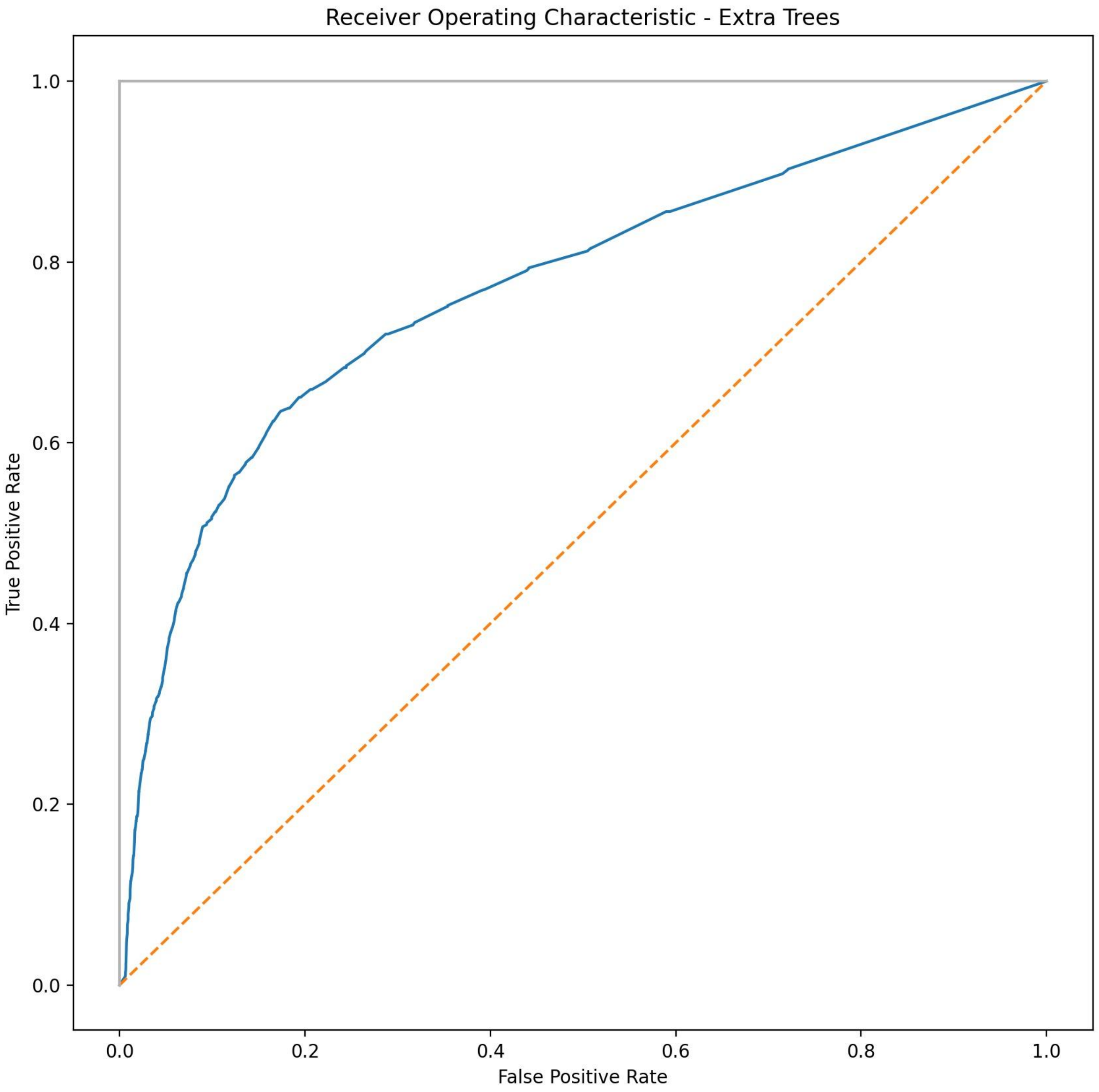


Stochastic Gradient Boosting

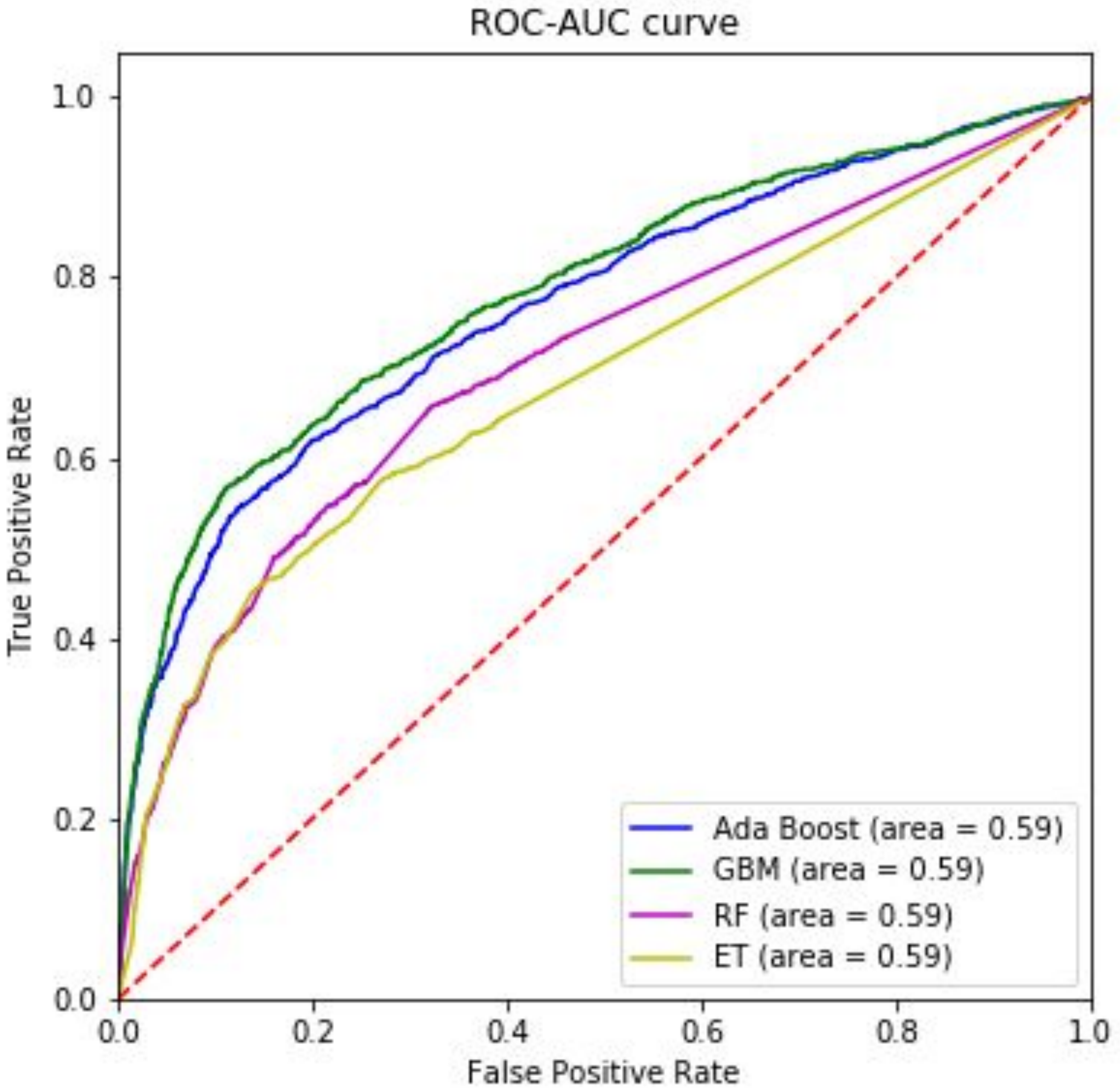


Random Forest





Extra Trees



We can see that both boosting techniques provide strong accuracy scores in the high 70% and even reached 80% with default configurations. The GBM model is the best model compared to the other ones. Therefore we will consider that model for production.

Model Recommendation

What type of ML model to use ?

Stochastic Gradient Boosting is the best model that fits with our classification task. Therefore, we proceed to save this model for later use.

Thank You

Data Glacier

Your Deep Learning Partner