

**Final Project Report**

**Bank Marketing Campaign**

**- Data Science -**

**Table of Contents**

**Team member’s details……………………………………….…………..…….....3**

**Problem description……………………………………….……………...…….....4**

**Business understanding………………….………………………………...……...4**

**Data Intake Report…………………………………………………….………….5**

**Data Understanding……………………………………………………………….7**

**Data Types…………………………………………………………………………9**

**Data Problems……………………………………………………………………10**

**Data Transformation………………………………………………………….…12**

**Data Dependency………………………………………………………………...13**

**Model Building……………………………………………………………….…..14**

**Results…………………………………………………………………………….15**

**Final Recommendation……………………………………...…………………...19**

**GitHub Repo Link…………...………………………………………….…….....19**

**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 |

**Problem description:**

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. In order to achieve this task, they approached an Analytics company to automate this process of classification. The Analytics company has given responsibility to the **Data Science Enthusiasts** Team and has asked to come up with a ML model to shortlist customers whose chances of buying the product is higher, so that ABC’s marketing channel can focus only on those customers.

**Business understanding:**

There has been a revenue decline for an ABC bank and they would like to know what actions to take. After investigation, they found out that the root cause is that their clients are not depositing as frequently as before. Knowing that term deposits allow banks to hold onto a deposit for a specific amount of time, banks can invest in higher gain financial products to make a profit.

In addition, banks also hold better chances to persuade term deposit clients into buying other products such as funds or insurance to further increase their revenues. As a result, the ABC bank would like to identify existing clients that have higher chances to subscribe for a term deposit and focus marketing efforts on such clients. The classification goal is to predict if the client will subscribe to a term deposit or not.

**Data Intake report:**

**Name:** Bank Marketing Campaign - Data Science

**Report date:** August 7th, 2021

**Internship Batch:** LISUM01

**Version:** 1.0

**Data intake by:** Data Science Enthusiasts Team

**Data intake reviewer:** Vatsal Vinesh Mandalia

**Data storage location:** <https://github.com/Vatsal-2414/Bank-Marketing-Campaign-DSEnthusiasts2021.git>

**Tabular data details: ‘bank.csv’**

|  |  |
| --- | --- |
| **Total number of observations** | 4521 |
| **Total number of files** | 1 |
| **Total number of features** | 17 |
| **Base format of the file** | .csv |
| **Size of the data** | 461 KB |

**Tabular data details: ‘bank-full.csv’**

|  |  |
| --- | --- |
| **Total number of observations** | 45211 |
| **Total number of files** | 1 |
| **Total number of features** | 17 |
| **Base format of the file** | .csv |
| **Size of the data** | 4.61 MB |

**Tabular data details: ‘bank-additional.csv’**

|  |  |
| --- | --- |
| **Total number of observations** | 4119 |
| **Total number of files** | 1 |
| **Total number of features** | 21 |
| **Base format of the file** | .csv |
| **Size of the data** | 584 KB |

**Tabular data details: ‘bank-additional-full.csv’**

|  |  |
| --- | --- |
| **Total number of observations** | 41118 |
| **Total number of files** | 1 |
| **Total number of features** | 21 |
| **Base format of the file** | .csv |
| **Size of the data** | 4.61 MB |

**Proposed Approach:**

* There are 12 rows of duplicated data in the ‘bank-additional-full’ dataset.
* There are no missing values in all datasets.

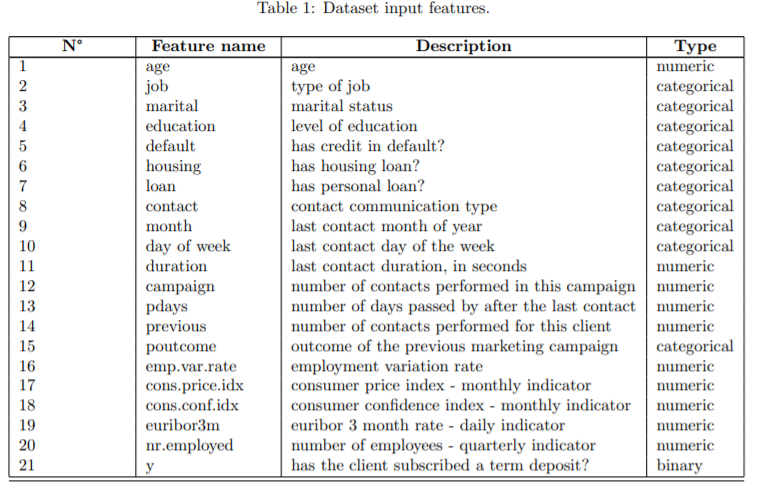
**Data Understanding:**

The data corresponds to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets provided for this classification problem. Among the four datasets, there are two pairs of train and test data available for analysis. The ‘bank-full.csv’ and ‘bank.csv’ are one of the pairs having less than 20 input features and are an older version of ‘bank-additional-full.csv’ and ‘bank-additional.csv’. The information of the datasets is given below.

|  |  |  |
| --- | --- | --- |
|  | Dataset type | Description |
| bank-additional-full.csv | train | 41118 observations and 20 inputs ordered by date (from May 2008 to November 2010) |
| bank-additional.csv | test | 4118 observations (10% of train data) with 20 inputs |
| bank-full.csv | train | 45211 observations and 17 inputs ordered by date (older version of bank-additional-full) |
| bank.csv | test | 4521 observations (10% of train data) and 17 inputs |

**Data set features description:**

****

The first 7 features are Bank client data.

Features 8, 9, 10 and 11 are related to the last contact of the current campaign.

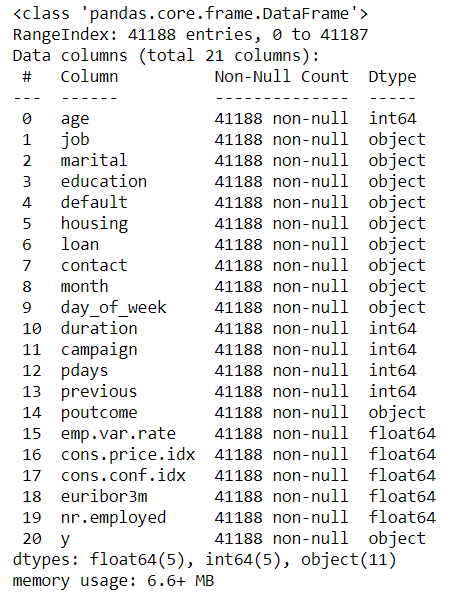
Features 12, 13, 14 and 15 are other attributes.

Features 16, 17, 18, 19 and 20 are social and economic context attributes.

The last feature is the output feature (desired target).

**Data Types**

In this dataset the features that we described above, are divided between “object” types i.e. categorical attributes and “int64 / float64” types i.e. numerical attributes.

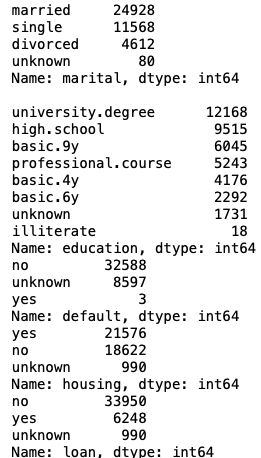
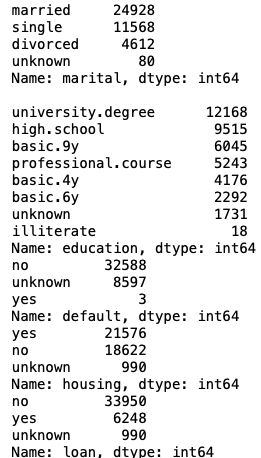
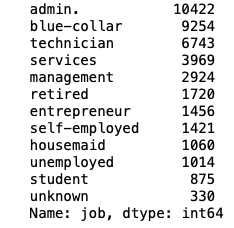


**Data Problems:**

**Missing Attribute Values:**

All 4 datasets have no missing values.

**Note:** There are a significant number of observations/rows with a value ‘unknown’ for the majority of the categorical features in the four datasets. So we assume the value ‘unknown’ as another category for these variables in our analysis.



**Duplicate rows:**

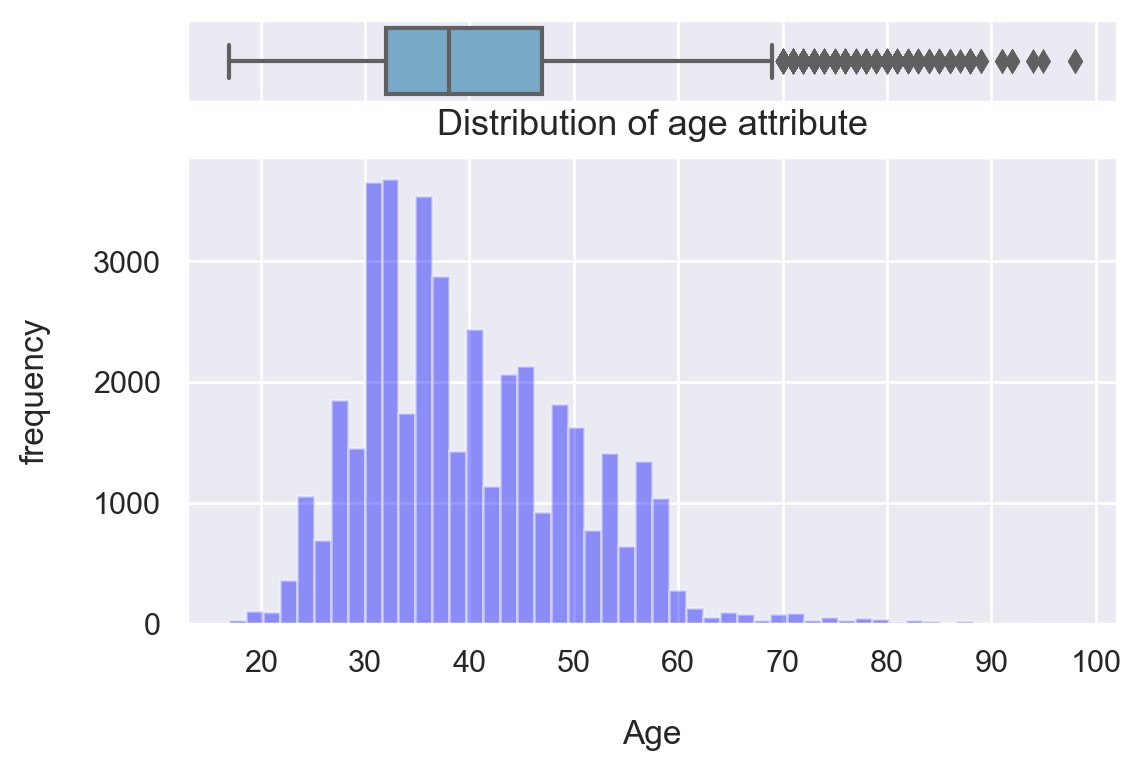
There are 12 rows whose duplicates are present in bank-additional-full data. The .drop\_duplicates() method is used to drop the duplicate rows.

In bank-additional, bank-full and bank.csv datasets, there are no duplicate rows.

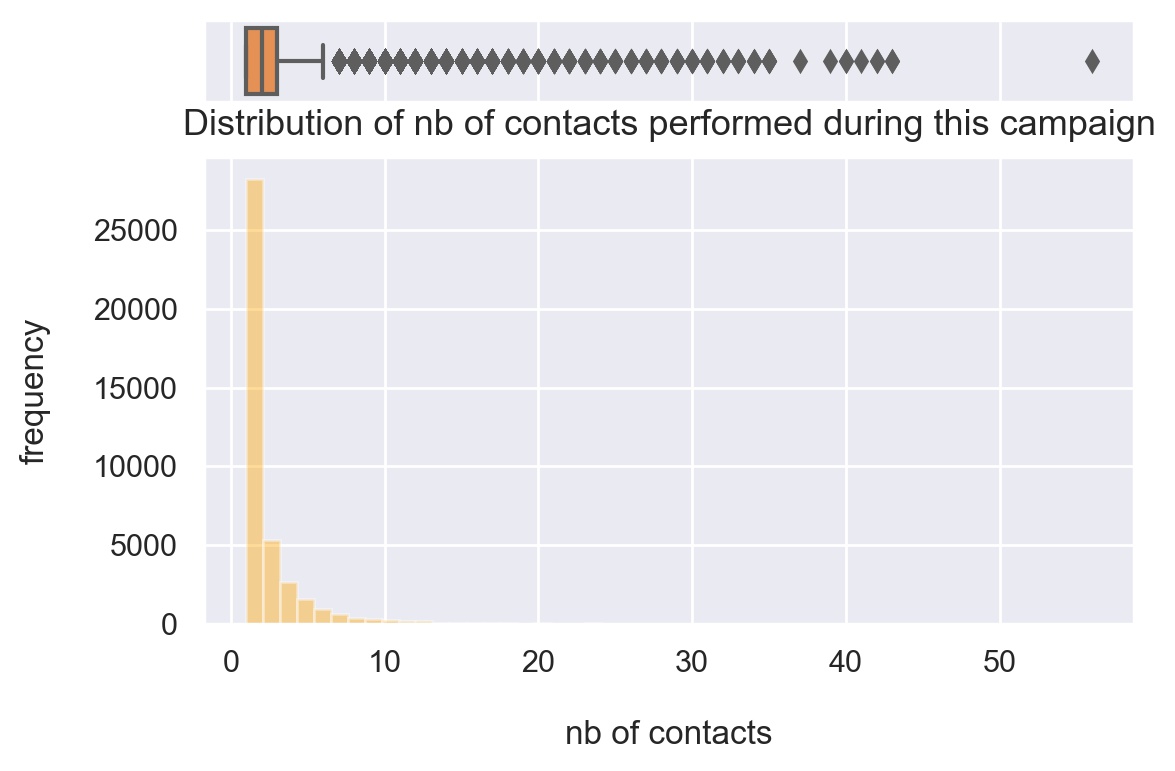
**Outliers:**

An outlier is a value/observation which lies at an abnormal distance from other values in the normal distribution. It can occur due to an error in measurement or data collection. The following features of bank-additional-full have significant outliers. Outliers can affect the mean of the distribution.

* **‘age’** attribute:



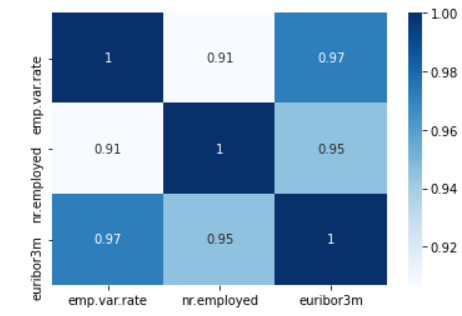
* **‘campaign’** attribute:



In our case, we don’t need to remove outliers from the data since the max(‘age’)=98 and max(‘campaign’)=56 are not unrealistic values. This will help with the generalization of the model later since it should reflect the real world.

**Data Transformation:**

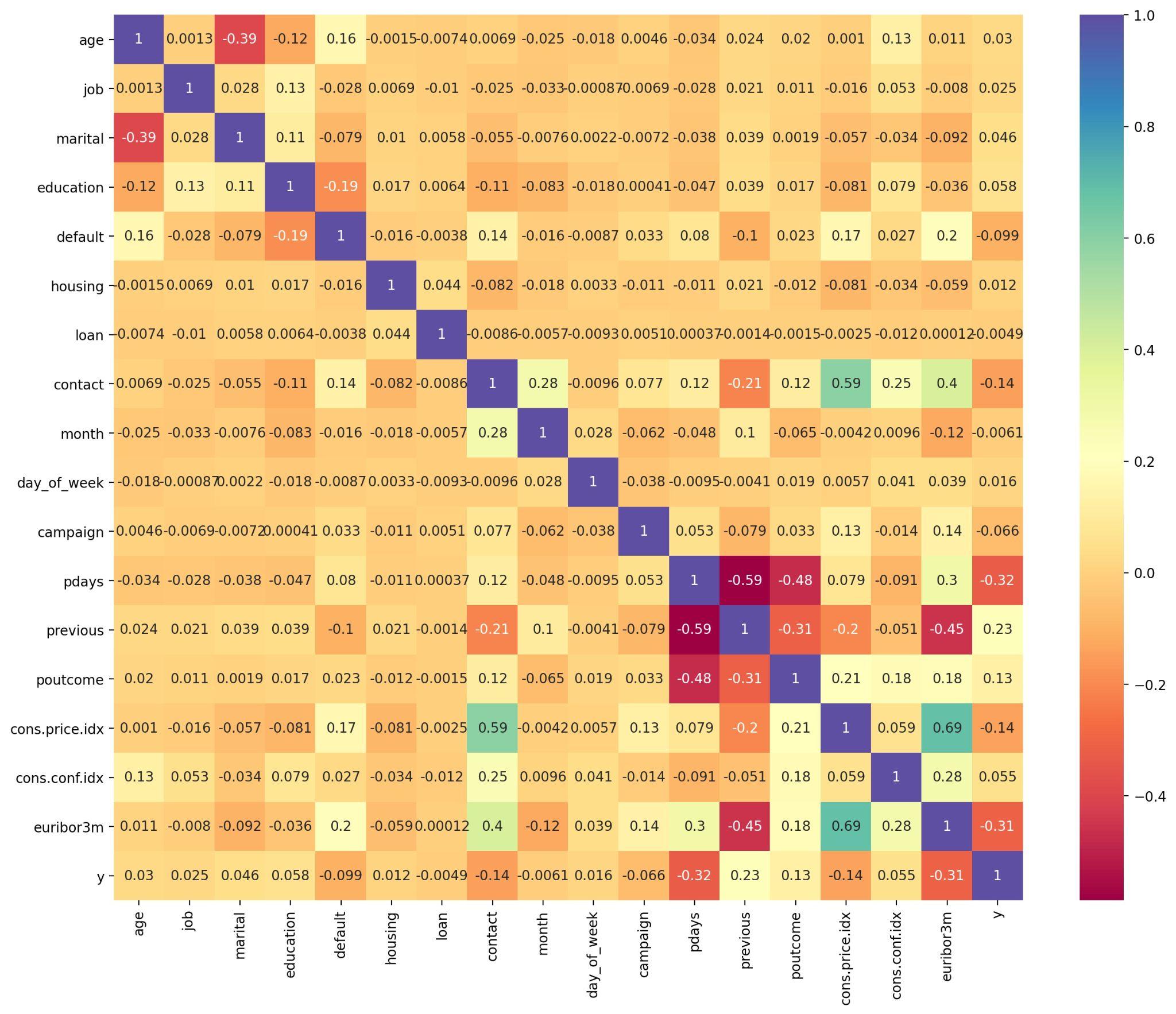
* Drop 'Duration' feature: This attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
* Remove duplicated rows.
* 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' and also 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. If we had, we could have written the condition that if the correlation is higher than 0.8 (or can be any threshold value depending on the domain knowledge) and less than -0.8, we could have dropped those features.



**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),

Gaussian Naive Bayes (NB)

k-Nearest Neighbors (KNN).

* **Ensemble Methods:**

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

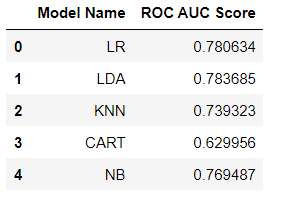
**Bagging Methods:** Random Forests (RF) and Extra Trees (ET).

**Results from Modelling:**

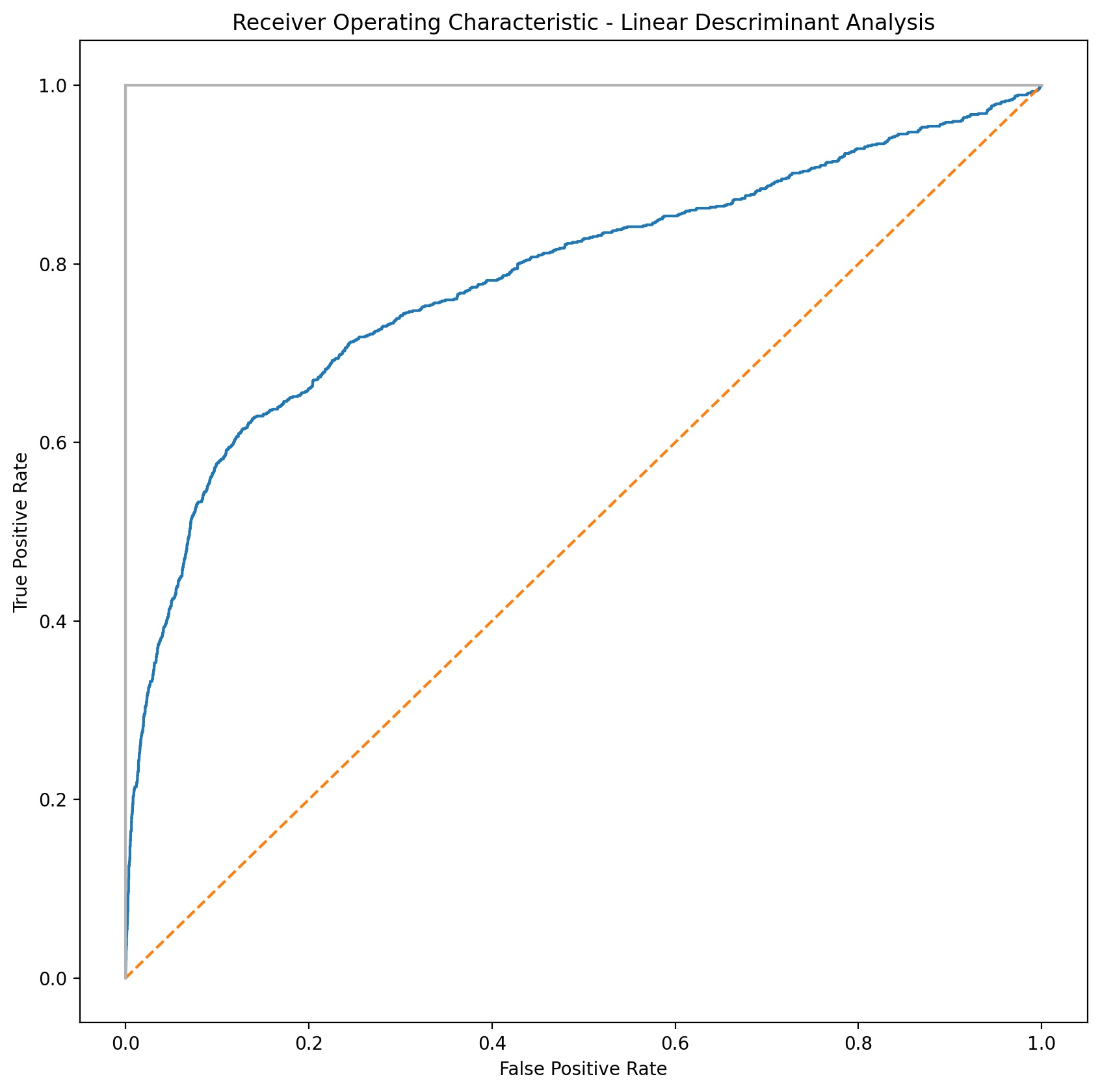
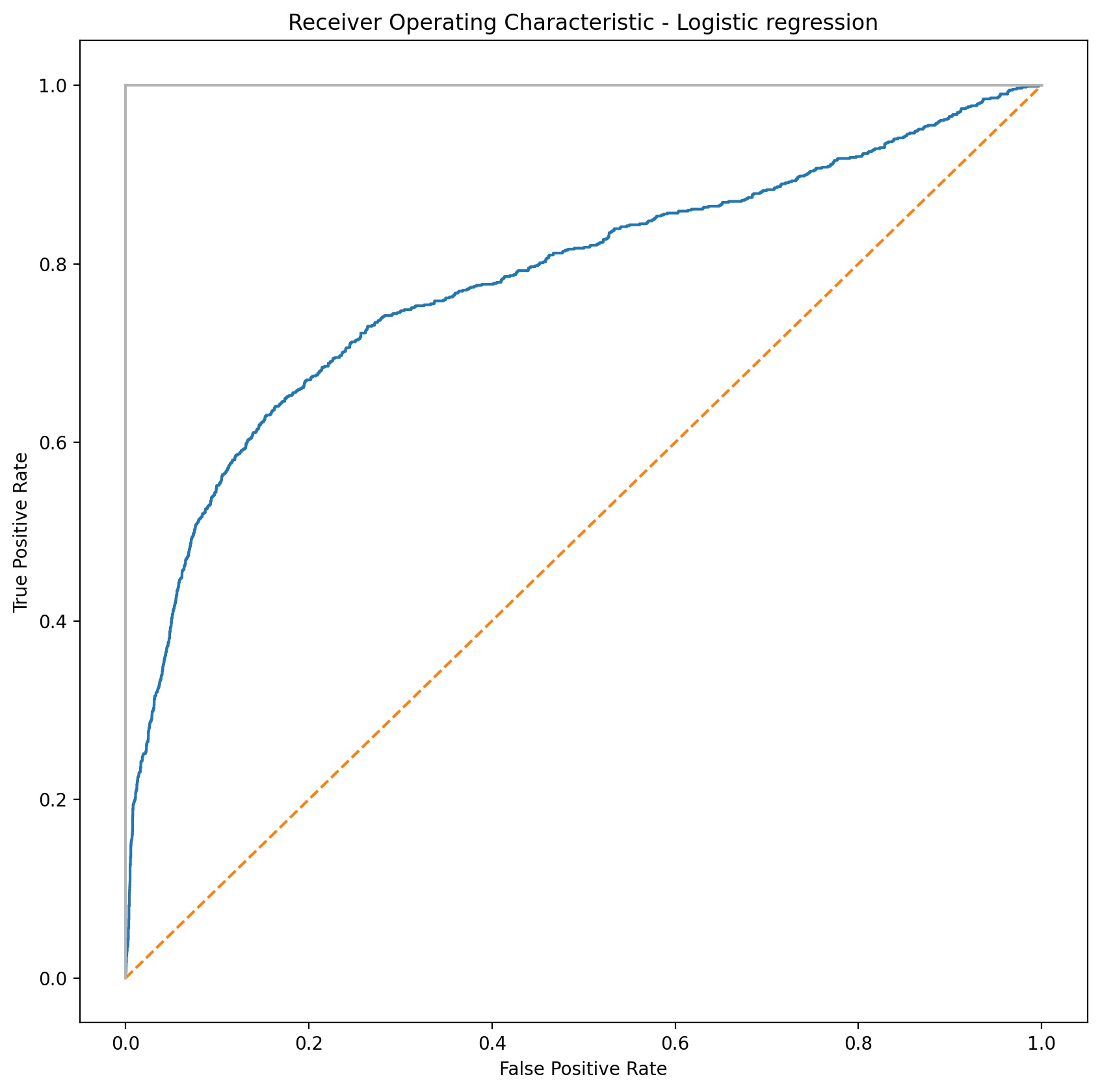
We suspected that the differing distributions of the raw data may be negatively impacting the skill of some of the algorithms. Therefore, we evaluated the same algorithms with a standardized copy of the dataset.

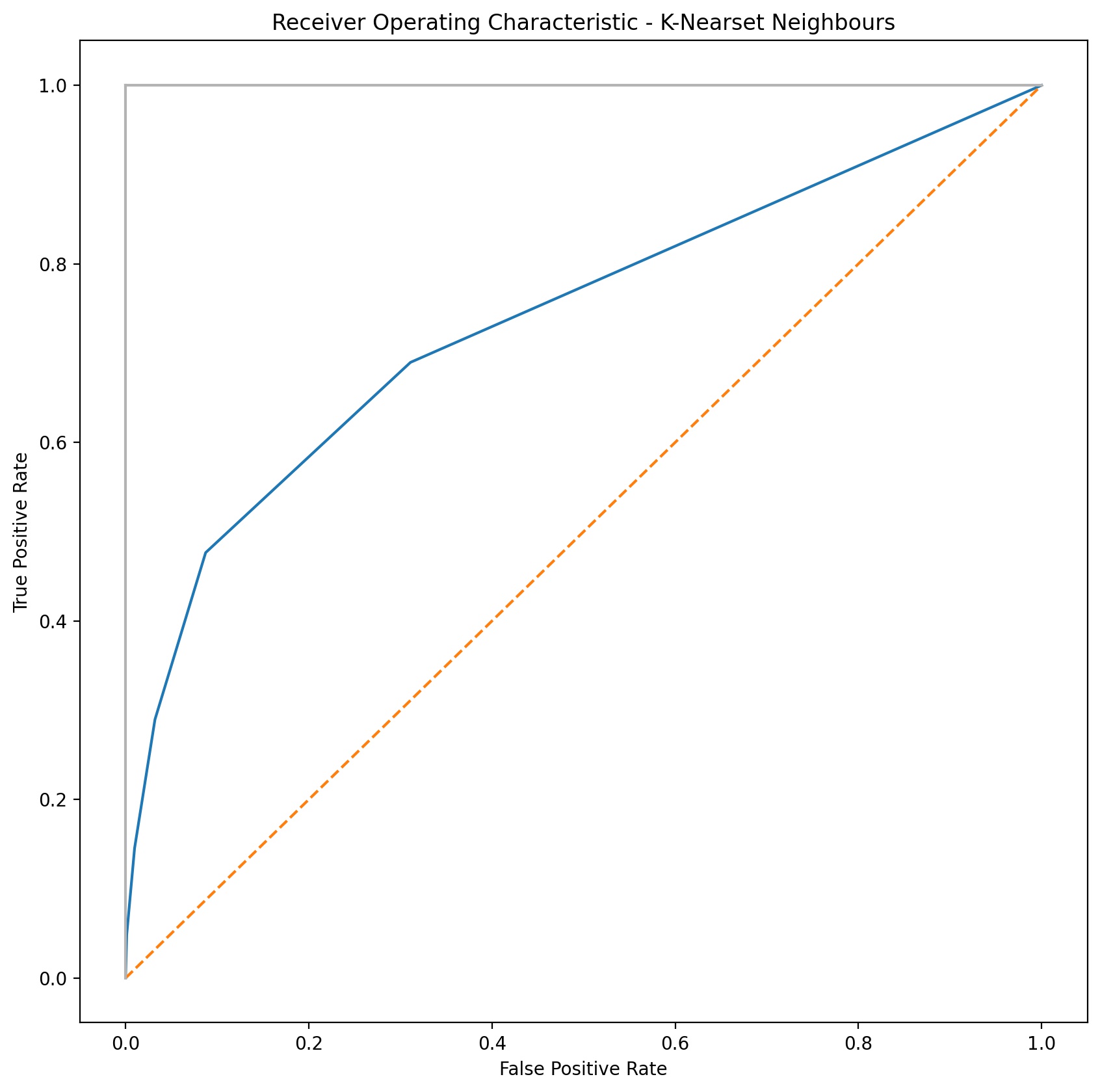
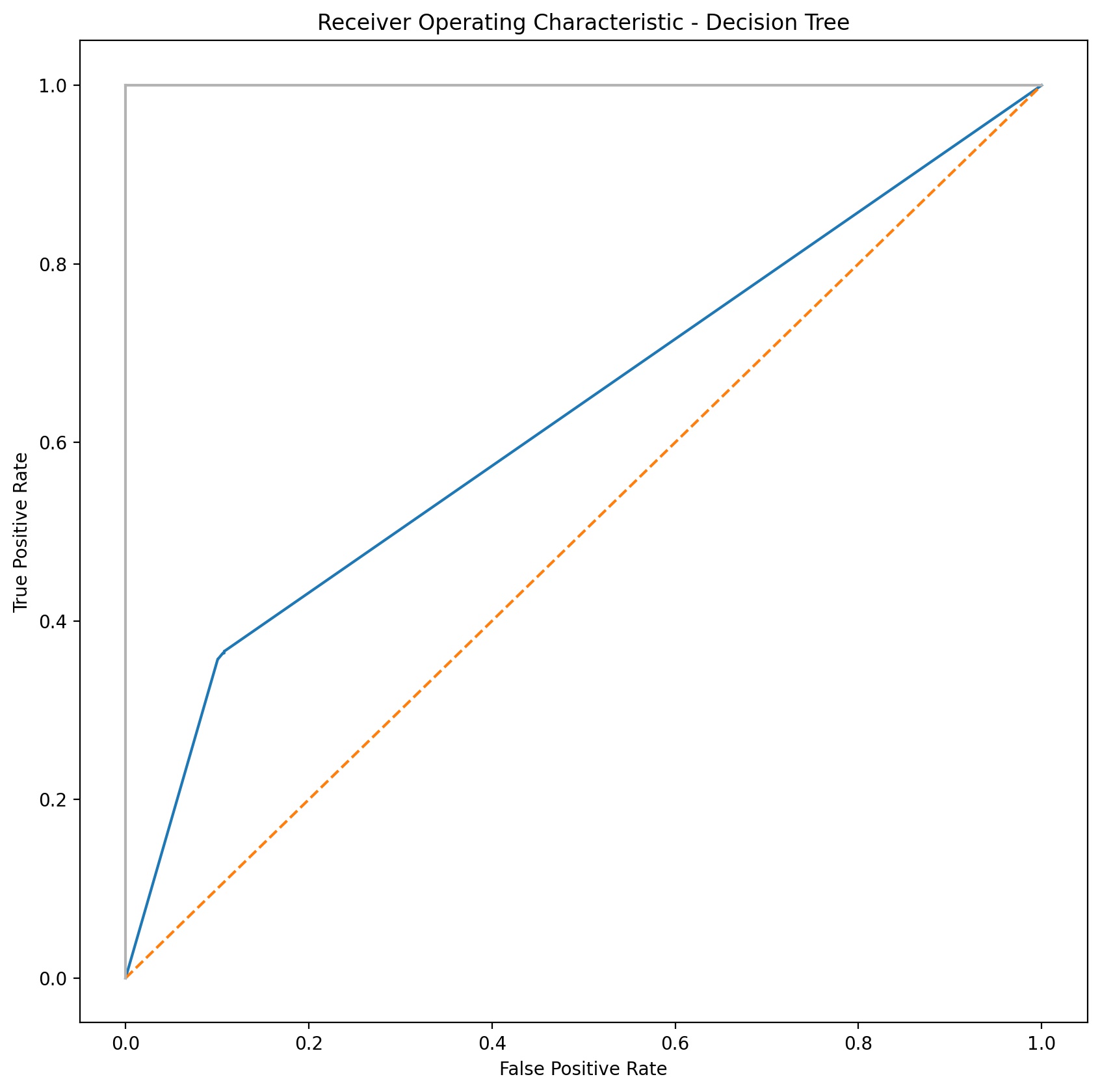
This is where the data is transformed such that each attribute has a mean value of zero and a standard deviation of one. We also need to avoid data leakage when we transform the data. A good way to avoid leakage is to use pipelines that standardize the data and build the model for each fold in the cross validation test harness. That way we can get a fair estimation of how each model with standardized data might perform on unseen data.

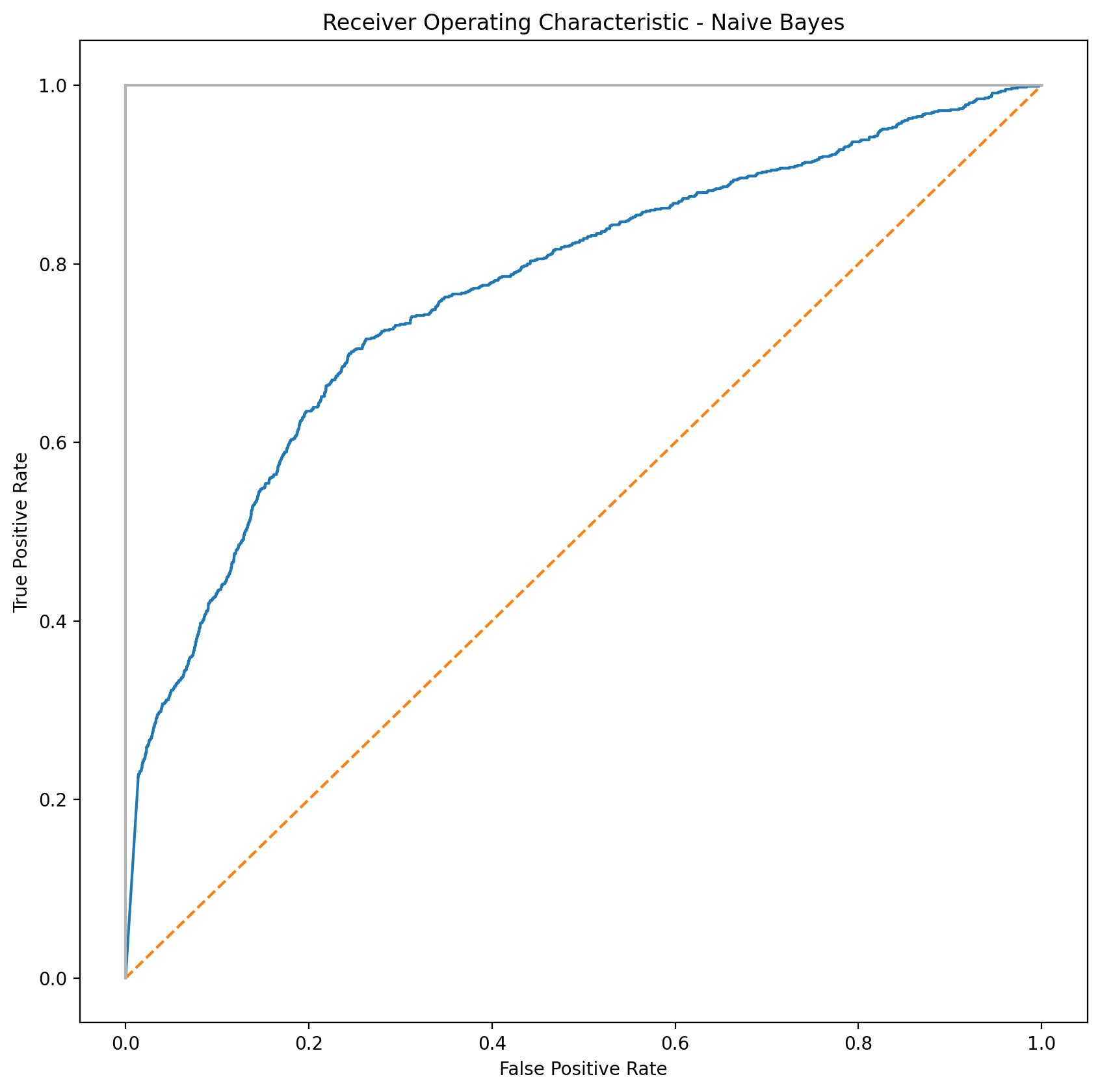
**Linear Algorithms & Nonlinear Algorithms results:**

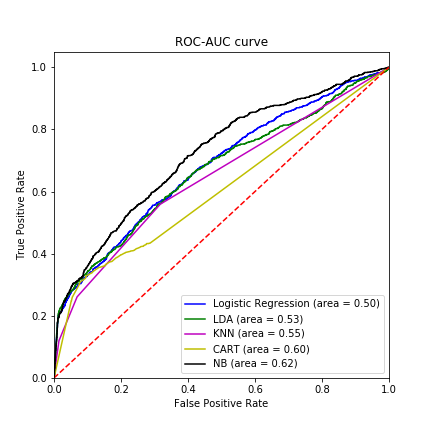


Logistic Regression is used as a base model in our case. As shown in the table, LDA gave better results than LR. Here are the ROC AUC Curves plotted separately for each model.

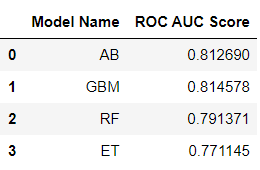
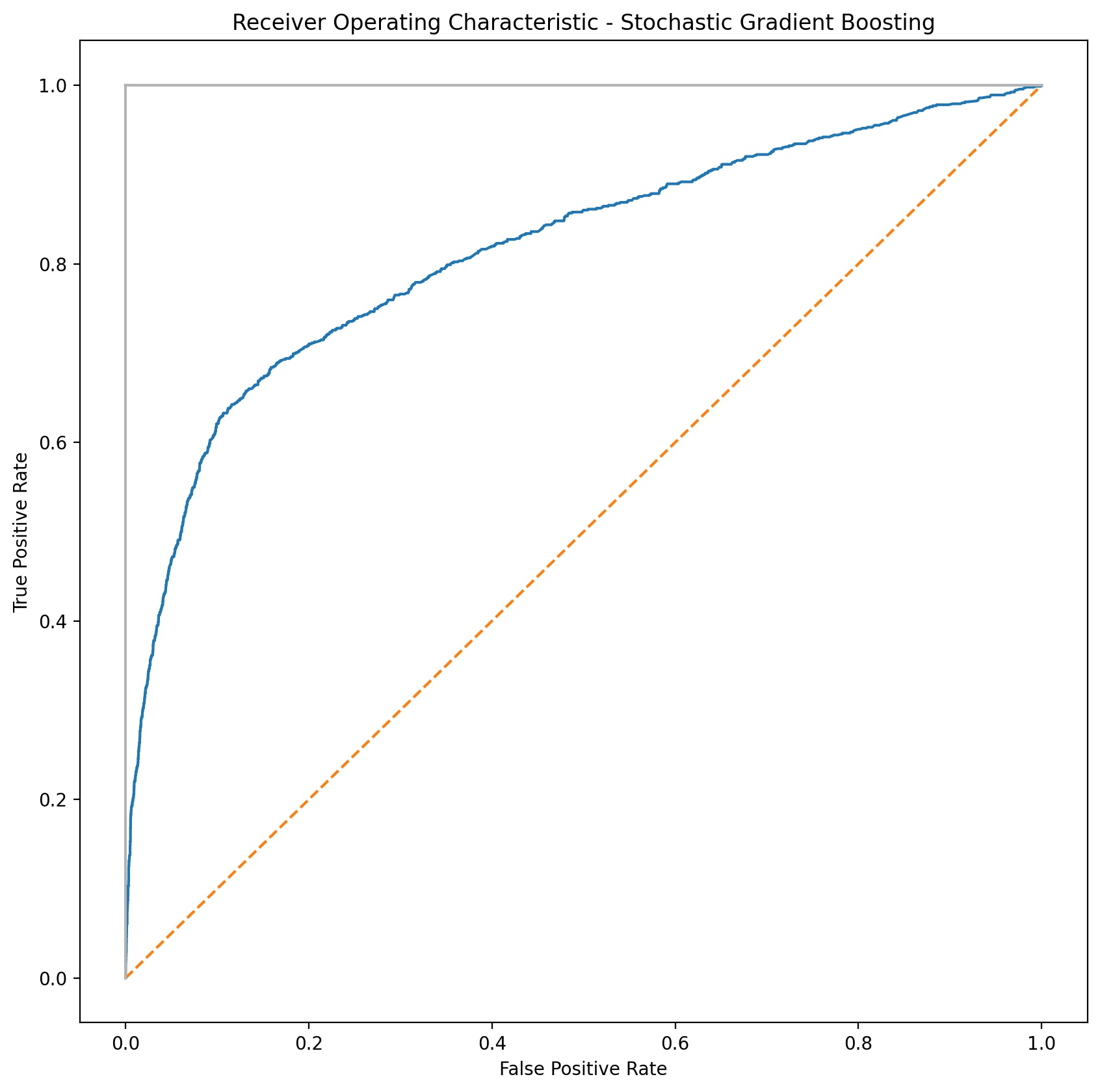
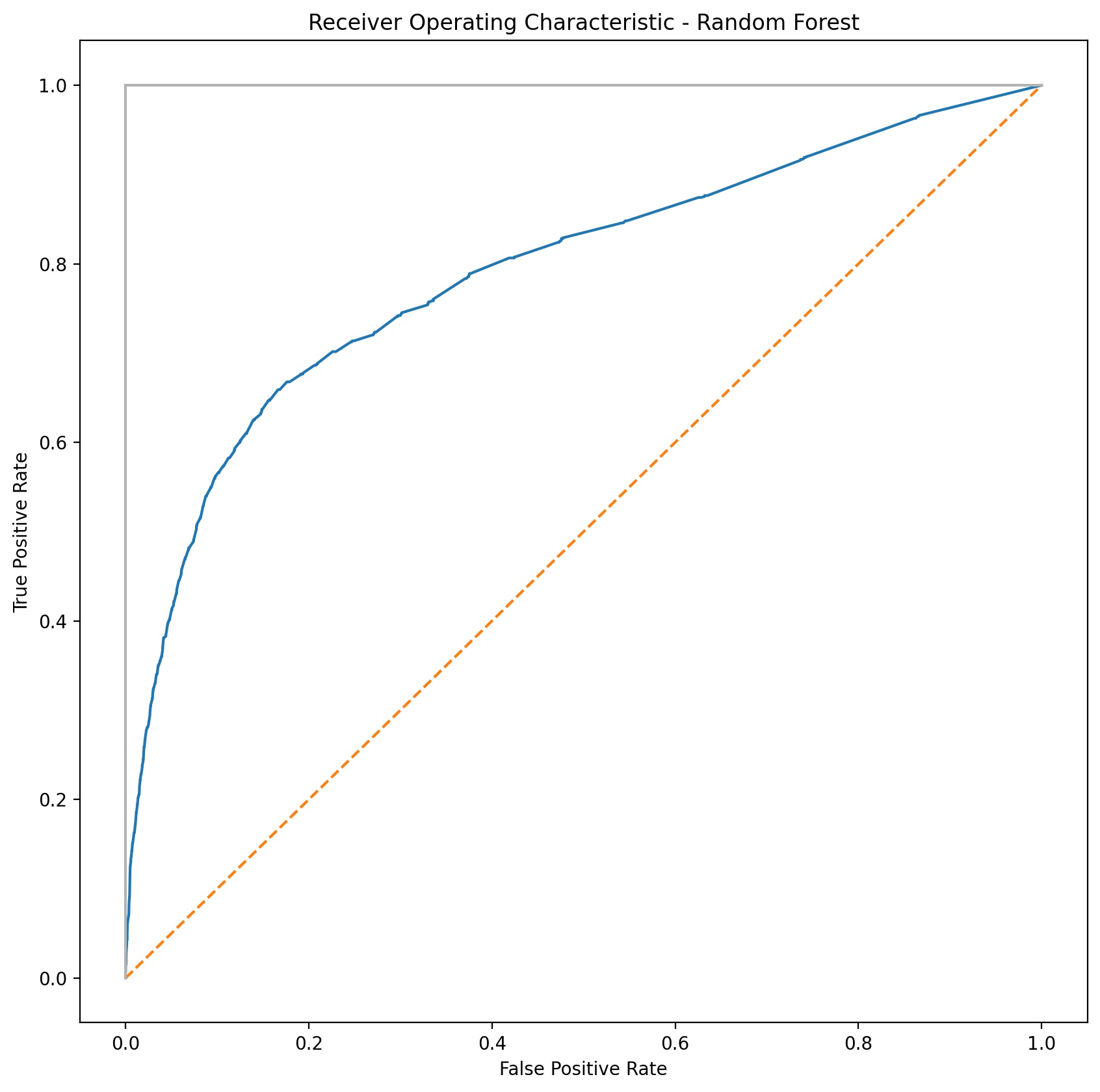
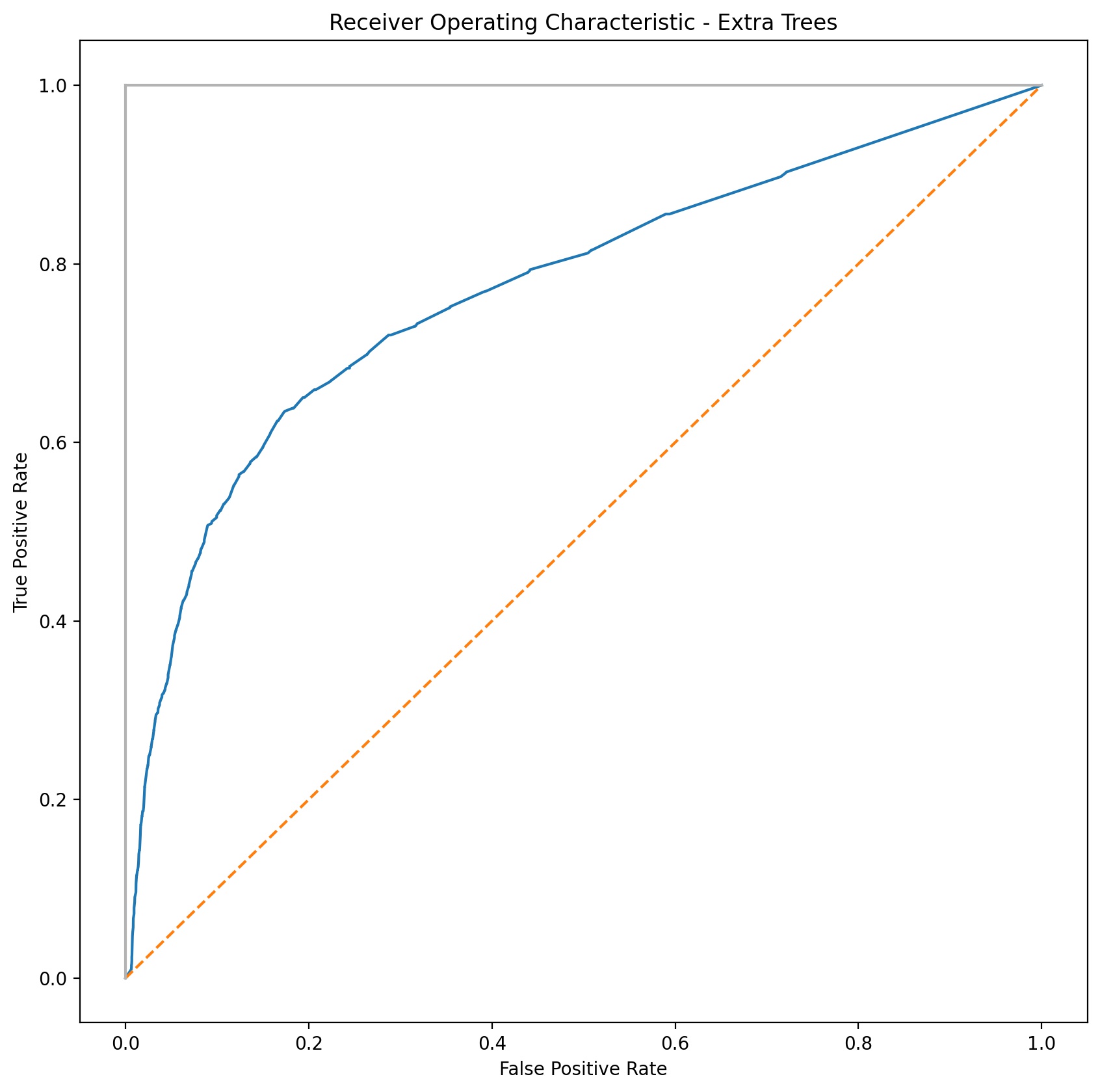
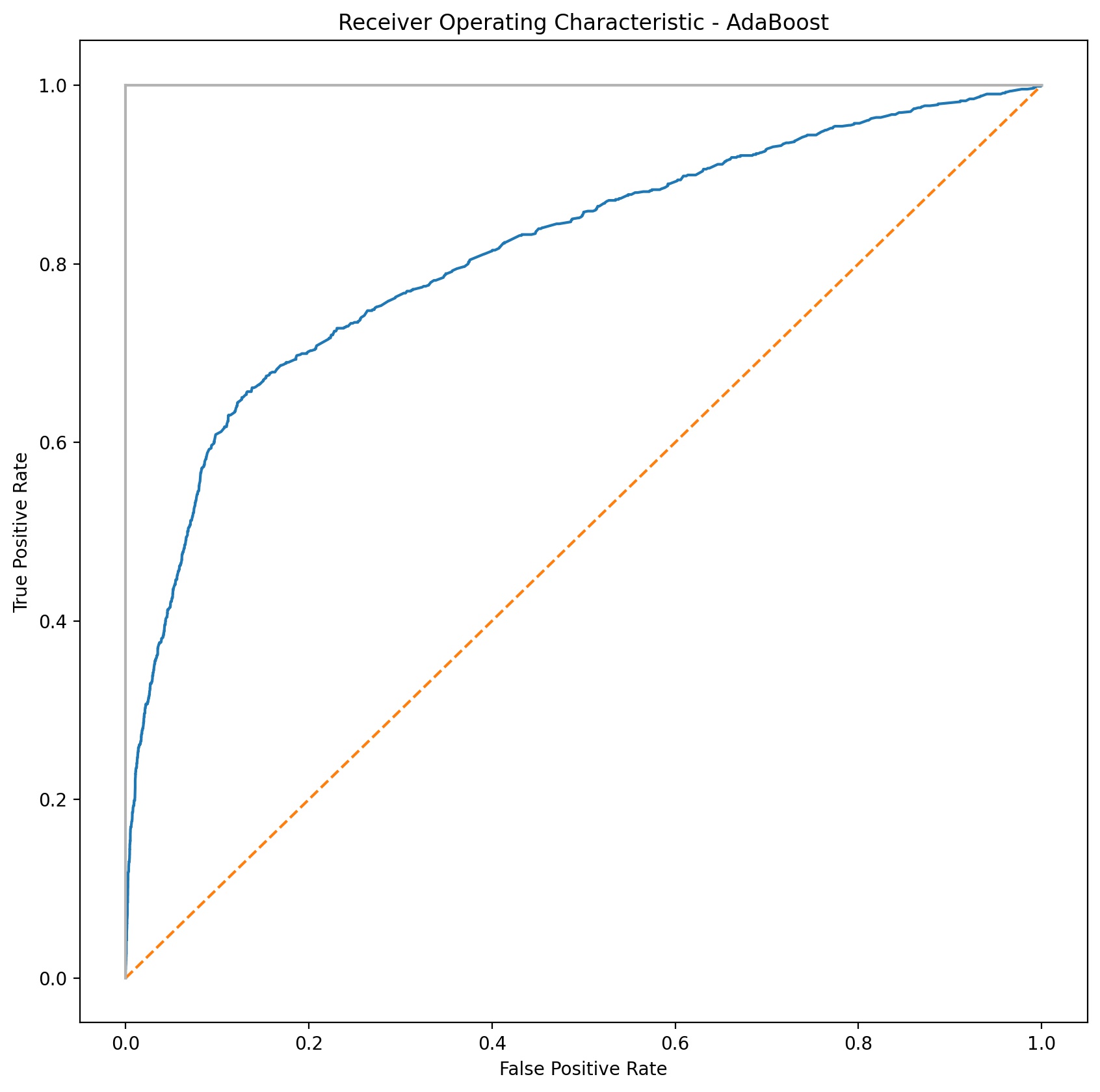


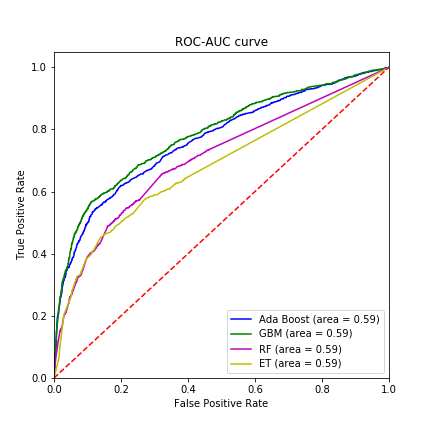




**Figure: ROC\_AUC curves for the linear and non-linear algorithms. For the plot, Standard Scaler and Cross Validation is not used to get a basic result.**

**Ensemble Methods results:**

****

****

**Figure: ROC\_AUC curves for Ensemble Methods Algorithms. Curves produced from data without using Standard Scaler and Cross Validation to get a basic result.**

**Final recommendation:**

We can see that both boosting techniques provide strong accuracy scores in the high 70s (%) 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.

**Github Repo link:**

<https://github.com/Vatsal-2414/Bank-Marketing-Campaign-DSEnthusiasts2021.git>

This repository includes the four datasets, model code (filename : Final\_Week\_code.ipynb) and necessary files used in this project.