Homework_02_WI25

February 4, 2025

1 Problem 1

Radioactive decay of a nucleus is described as $N(t) = N_0 e^{-t/\tau}$ where N_0 is the number of the nucleus in question initially and τ is the half life of the nucleus. $N(t)$ is the number of nuclei remaining at time t.

1.1 Part 1

propagate the uncertainties on N_0 and τ to get the corresponding uncertainty on $N(t)$ analytically using the information discussed in class.

Type your answer below:

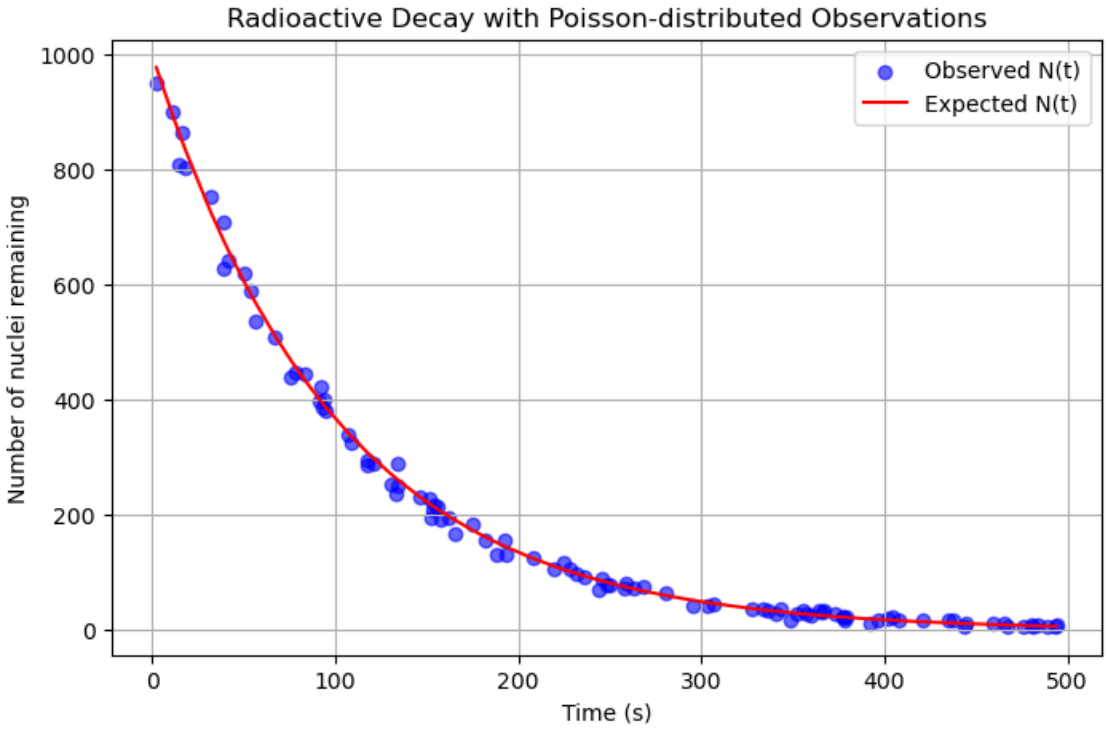
$\sigma_{N(t)} = e^{-\frac{t}{\tau}} \sqrt{\left(\sigma_{N_0}\right)^2 + \left(\frac{N_0 t}{\tau^2} \cdot \sigma_\tau\right)^2}$

1.2 Part 2

Write a program that draws random numbers according to what you would expect to measure in terms of $N(t)$ as a function of time for a given τ. Set N_0 to at least 1k, generate 100 samples for uniform random t-values in a reasonable range; samples at larger times will of course be more rare. For the plot purposes set $\tau = 100s$. Make a plot out to large enough t values.

[1]:
```
import numpy as np
import matplotlib.pyplot as plt

N0 = 1000
tau = 100
num_samples = 100
t_max = 500

t_values = np.random.uniform(0, t_max, num_samples)
N_expected = N0 * np.exp(-t_values / tau)
N_observed = np.random.poisson(N_expected)

sorted_indices = np.argsort(t_values)
t_values_sorted = t_values[sorted_indices]
N_observed_sorted = N_observed[sorted_indices]
```

[2]:
```
plt.figure(figsize=(8, 5))
plt.scatter(t_values_sorted, N_observed_sorted, color='blue', alpha=0.6,␣
 ↪label="Observed N(t)")
plt.plot(t_values_sorted, N0 * np.exp(-t_values_sorted / tau), 'r-',␣
 ↪label="Expected N(t)")
plt.xlabel("Time (s)")
plt.ylabel("Number of nuclei remaining")
plt.title("Radioactive Decay with Poisson-distributed Observations")
plt.legend()
plt.grid(True)
plt.show()
```



1.3 Part 3 (Bonus)

Fit $N(t)$ using a least squares fit to get τ and N_0. Make sure to keep track of the statistical uncertainties in your *data* and your *fit*. Do the values agree with those you set in *Part 2*? Are τ and $N(0)$ correlated with each other?

2 Problem 2: Tabular data and BDTs: Classifying LHC collisions

2.0.1 Goal

Discriminate $H \to \tau\tau$ (signal) from background such as $t\bar{t}$.

In a real detector, the signal looks like:

2.0.2 Boosted decision tree library: XGBoost

We'll use the python API for the XGBoost (eXtreme Gradient Boosting) library.

2.0.3 Data

ATLAS hosted a Kaggle competition for identifying $H \to \tau\tau$ events, the Higgs Boson Machine Learning Challenge. The training data for this event contains 250,000 labeled, simulated ATLAS events in `csv` format described here and here. You can download it yourself, but we will only play with a small subset (10k events).

2.0.4 Data handling

We'll use Pandas.

2.0.5 Installing XGBoost

Assuming you have Python, NumPy, Matplotlib, and Pandas installed, you may need to install XGBoost if it's not already installed.

```
pip install xgboost --user
```

2.0.6 Links

A lot of this was borrowed from other sources. These sources and other good places for information about XGBoost and BDTs in general are here: * XGBoost demo: Example of how to use XGBoost Python module to run Kaggle Higgs competition * Blog post by phunther: Winning solution of Kaggle Higgs competition: what a single model can do? * XGBoost Kaggle Higgs solution: https://github.com/hetong007/higgsml

2.0.7 Note

You will not be asked to do the model training yourselves. The train/test datasets will be processed and provided, with the XGBoost model trained ready to use. The main purpose of this problem is to get real hands-on practice of evaluating a model, like ROC, recall, etc.

2.1 XGBoost Tutorial

[2]:
```
!pip install xgboost
```

```
Collecting xgboost
  Obtaining dependency information for xgboost from https://files.pythonhosted.o
rg/packages/32/93/66826e2f50cefecbb0a44bd1e667316bf0a3c8e78cd1f0cdf52f5b2c5c6f/x
gboost-2.1.3-py3-none-manylinux_2_28_x86_64.whl.metadata
  Downloading xgboost-2.1.3-py3-none-manylinux_2_28_x86_64.whl.metadata (2.1 kB)
Requirement already satisfied: numpy in /opt/conda/lib/python3.11/site-packages
(from xgboost) (1.24.3)
Collecting nvidia-nccl-cu12 (from xgboost)
```

```
  Obtaining dependency information for nvidia-nccl-cu12 from https://files.pytho
nhosted.org/packages/11/0c/8c78b7603f4e685624a3ea944940f1e75f36d71bd6504330511f4
a0e1557/nvidia_nccl_cu12-2.25.1-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata
  Downloading nvidia_nccl_cu12-2.25.1-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.8 kB)
Requirement already satisfied: scipy in /opt/conda/lib/python3.11/site-packages
(from xgboost) (1.11.2)
Downloading xgboost-2.1.3-py3-none-manylinux_2_28_x86_64.whl (153.9 MB)
                         153.9/153.9 MB
28.2 MB/s eta 0:00:0000:0100:01
Downloading nvidia_nccl_cu12-2.25.1-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (201.4 MB)
                         201.4/201.4 MB
22.6 MB/s eta 0:00:0000:0100:01
Installing collected packages: nvidia-nccl-cu12, xgboost
Successfully installed nvidia-nccl-cu12-2.25.1 xgboost-2.1.3
```

[3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import xgboost as xgb

%matplotlib inline
```

2.1.1 Data

Load data First, load in the data and look at it. We will download a 10k event subsample of the Kaggle training data. Then we'll put it in the right format for XGBoost.

[6]:
```
!pwd
```

```
/j-jepa-vol/J-JEPA-Zihan/DSC_Homework
```

[8]:
```
!mkdir -p data
! wget https://raw.githubusercontent.com/k-woodruff/bdt-tutorial/master/data/
 ↪training_10k.csv -O data/training_10k.csv
```

```
--2025-01-27 22:10:48--  https://raw.githubusercontent.com/k-woodruff/bdt-
tutorial/master/data/training_10k.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)…
185.199.108.133, 185.199.111.133, 185.199.109.133, …
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.108.133|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 2210243 (2.1M) [text/plain]
Saving to: 'data/training_10k.csv'
```

```
data/training_10k.c 100%[===================>]   2.11M  --.-KB/s    in 0.02s

2025-01-27 22:10:48 (124 MB/s) - ‘data/training_10k.csv’ saved [2210243/2210243]
```

[9]:
```
data = pd.read_csv("data/training_10k.csv")
```

Let’s see what the data looks like:

[10]:
```
print("Size of data: {}".format(data.shape))
print("Number of events: {}".format(data.shape[0]))
print("Number of columns: {}".format(data.shape[1]))

print("\nList of features in dataset:")
for col in data.columns:
    print(col)
```

```
Size of data: (10000, 33)
Number of events: 10000
Number of columns: 33

List of features in dataset:
EventId
DER_mass_MMC
DER_mass_transverse_met_lep
DER_mass_vis
DER_pt_h
DER_deltaeta_jet_jet
DER_mass_jet_jet
DER_prodeta_jet_jet
DER_deltar_tau_lep
DER_pt_tot
DER_sum_pt
DER_pt_ratio_lep_tau
DER_met_phi_centrality
DER_lep_eta_centrality
PRI_tau_pt
PRI_tau_eta
PRI_tau_phi
PRI_lep_pt
PRI_lep_eta
PRI_lep_phi
PRI_met
PRI_met_phi
PRI_met_sumet
PRI_jet_num
PRI_jet_leading_pt
PRI_jet_leading_eta
```

PRI_jet_leading_phi
PRI_jet_subleading_pt
PRI_jet_subleading_eta
PRI_jet_subleading_phi
PRI_jet_all_pt
Weight
Label

2.1.2 Detailed description of features

Prefix-less variables EventId, Weight, and Label have a special role and should not be used as input to the classifier. The variables prefixed with PRI (for PRImitives) are "raw" quantities about the bunch collision as measured by the detector, essentially the momenta of particles. Variables prefixed with DER (for DERived) are quantities computed from the primitive features. These quantities were selected by the physicists of ATLAS in the reference document either to select regions of interest or as features for the Boosted Decision Trees used in this analysis. In addition: * Variables are floating point unless specified otherwise. * All azimuthal ϕ angles are in radian in the $[-\pi, +\pi]$ range. * Energy, mass, momentum are all in GeV * All other variables are unitless. * Variables are indicated as "may be undefined" when it can happen that they are meaningless or cannot be computed; in this case, their value is -999.0, which is outside the normal range of all variables. * The mass of particles has not been provided, as it can safely be neglected for the Challenge.

Features: - EventId: An unique integer identifier of the event. Not to be used as a feature. - DER_mass_MMC: The estimated mass m_H of the Higgs boson candidate, obtained through a probabilistic phase space integration (may be undefined if the topology of the event is too far from the expected topology) - DER_mass_transverse_met_lep: The transverse mass (21) between the missing transverse energy and the lepton. - DER_mass_vis: The invariant mass (20) of the hadronic tau and the lepton. - DER_pt_h: The modulus (19) of the vector sum of the transverse momentum of the hadronic tau, the lepton, and the missing transverse energy vector. - DER_deltaeta_jet_jet: The absolute value of the pseudorapidity separation (22) between the two jets (undefined if PRI_jet_num ≤ 1). - DER_mass_jet_jet: The invariant mass (20) of the two jets (undefined if PRI_jet_num ≤ 1). - DER_prodeta_jet_jet: The product of the pseudorapidities of the two jets (undefined if PRI_jet_num ≤ 1). - DER_deltar_tau_lep: The R separation (23) between the hadronic tau and the lepton. - DER_pt_tot: The modulus (19) of the vector sum of the missing transverse momenta and the transverse momenta of the hadronic tau, the lepton, the leading jet (if PRI_jet_num ≥ 1) and the subleading jet (if PRI_jet_num $= 2$) (but not of any additional jets). - DER_sum_pt: The sum of the moduli (19) of the transverse momenta of the hadronic tau, the lepton, the leading jet (if PRI_jet_num ≥ 1) and the subleading jet (if PRI_jet_num $= 2$) and the other jets (if PRI_jet_num $= 3$). - DER_pt_ratio_lep_tau: The ratio of the transverse momenta of the lepton and the hadronic tau. - DER_met_phi_centrality: The centrality of the azimuthal angle of the missing transverse energy vector w.r.t. the hadronic tau and the lepton $C = \frac{A+B}{A^2+B^2}$ where $A = \sin(\phi_{\text{met}} - \phi_{\text{lep}})$, $B = \sin(\phi_{\text{had}} - \phi_{\text{met}})$, and ϕ_{met}, ϕ_{lep}, and ϕ_{had} are the azimuthal angles of the missing transverse energy vector, the lepton, and the hadronic tau, respectively. The centrality is $\sqrt{2}$ if the missing transverse energy vector $\vec{E}_{\text{T}}^{\text{miss}}$ is on the bisector of the transverse momenta of the lepton and the hadronic tau. It decreases to 1 if $\vec{E}_{\text{T}}^{\text{miss}}$ is collinear with one of these vectors and it decreases further to $-\sqrt{2}$ when $\vec{E}_{\text{T}}^{\text{miss}}$ is exactly opposite to the bisector. - DER_lep_eta_centrality: The centrality of the pseudorapidity of the lepton w.r.t. the

two jets (undefined if `PRI_jet_num` ≤ 1) $\exp\left[\frac{-4}{(\eta_1-\eta_2)^2}\left(\eta_{\text{lep}}-\frac{\eta_1+\eta_2}{2}\right)^2\right]$ where η_{lep} is the pseudorapidity of the lepton and η_1 and η_2 are the pseudorapidities of the two jets. The centrality is 1 when the lepton is on the bisector of the two jets, decreases to $1/e$ when it is collinear to one of the jets, and decreases further to zero at infinity. - `PRI_tau_pt`: The transverse momentum $\sqrt{p_x^2+p_y^2}$ of the hadronic tau. - `PRI_tau_eta`: The pseudorapidity η of the hadronic tau. - `PRI_tau_phi`: The azimuth angle ϕ of the hadronic tau. - `PRI_lep_pt`: The transverse momentum $\sqrt{p_x^2+p_y^2}$ of the lepton (electron or muon). - `PRI_lep_eta`: The pseudorapidity η of the lepton. - `PRI_ep_phi`: The azimuth angle ϕ of the lepton. - `PRI_met`: The missing transverse energy $E_{\text{T}}^{\text{miss}}$ - `PRI_met_phi` The azimuth angle ϕ of the missing transverse energy. - `PRI_met_sumet`: The total transverse energy in the detector. - `PRI_jet_num`: The number of jets (integer with value of 0, 1, 2 or 3; possible larger values have been capped at 3). - `PRI_jet_leading_pt`: The transverse momentum $\sqrt{p_x^2+p_y^2}$ of the leading jet, that is the jet with largest transverse momentum (undefined if `PRI_jet_num` $= 0$). - `PRI_jet_leading_eta`: The pseudorapidity η of the leading jet (undefined if `PRI_jet_num` $= 0$). - `PRI_jet_leading_phi`: The azimuth angle ϕ of the leading jet (undefined if `PRI_jet_num` $= 0$). - `PRI_jet_subleading_pt`: The transverse momentum $\sqrt{p_x^2+p_y^2}$ of the leading jet, that is, the jet with second largest transverse momentum (undefined if `PRI_jet_num` ≤ 1). - `PRI_jet_subleading_eta`: The pseudorapidity η of the subleading jet (undefined if `PRI_jet_num` ≤ 1). - `PRI_jet_subleading_phi`: The azimuth angle ϕ of the subleading jet (undefined if `PRI_jet_num` ≤ 1). - `PRI_jet_all_pt`: The scalar sum of the transverse momentum of all the jets of the events. - `Weight`: The event weight w_i, explained in Section 3.3. Not to be used as a feature. Not available in the test sample. - `Label`: The event label (string) $y_i \in \{s, b\}$ (s for signal, b for background). Not to be used as a feature. Not available in the test sample.

The data set has 10,000 events with 33 columns each. The first column is an identifier, and should not be used as a feature. The last two columns `Weight` and `Label`, are the weights and labels from the simulation, and also should not be used as features (this information is all contained in the documentation).

Now we can look at how many events are signal and background:

[11]:
```
# look at column labels --- notice last one is "Label" and first is "EventId"␣
  ↪also "Weight"
print(f"Number of signal events: {len(data[data.Label == 's'])}")
print(f"Number of background events: {len(data[data.Label == 'b'])}")
print(f"Fraction signal: {len(data[data.Label == 's'])/(len(data[data.Label ==␣
  ↪'s']) + len(data[data.Label == 'b']))}")
```

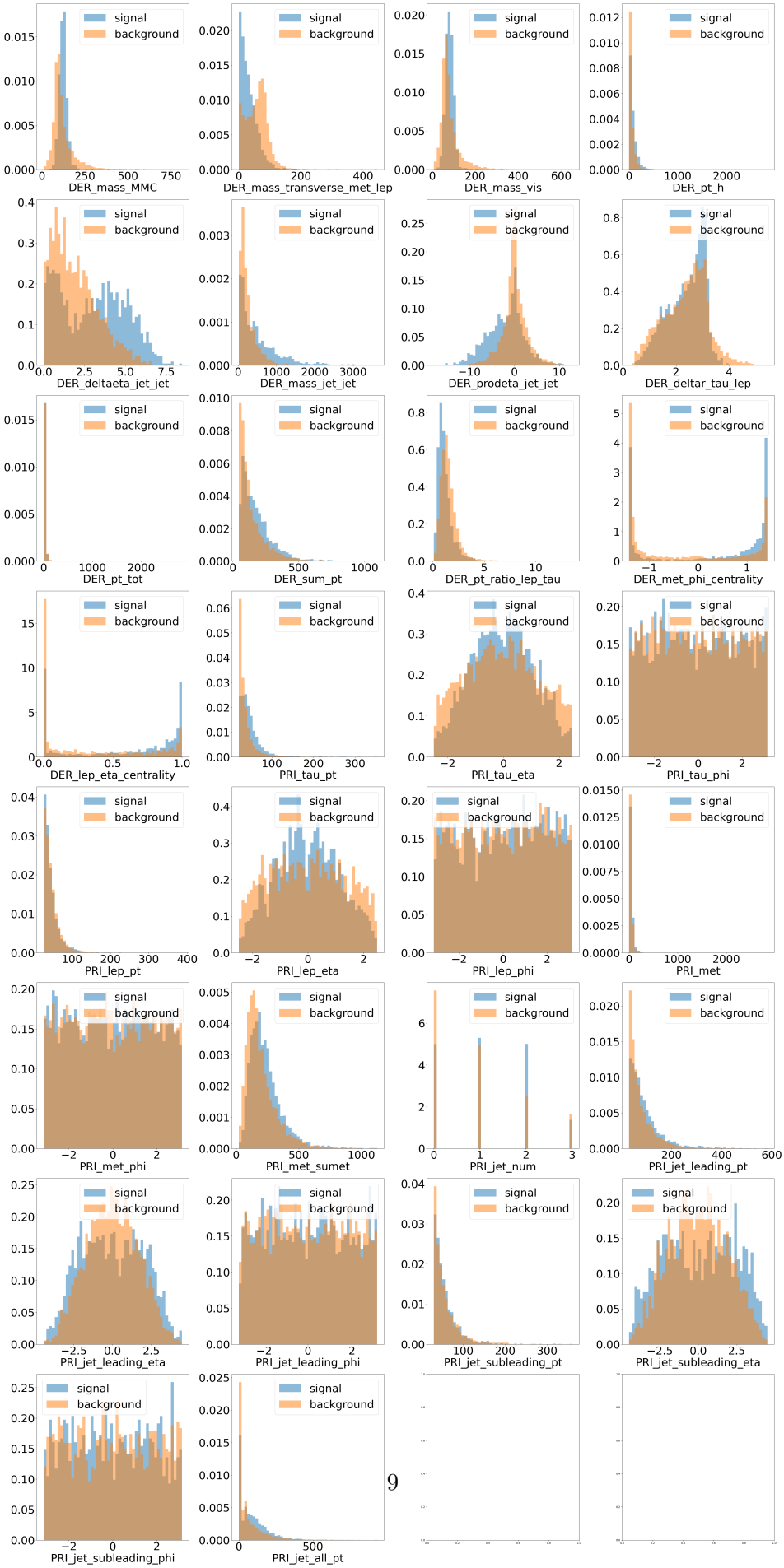
```
Number of signal events: 3372
Number of background events: 6628
Fraction signal: 0.3372
```

Visualize the features:

[12]:
```
plt.figure()
```

```
fig, axs = plt.subplots(8, 4, figsize=(40, 80))

for ix, ax in enumerate(axs.reshape(-1)):
    col = data.columns[ix + 1]
    if col == "Weight" or col == "Label":
        continue
    signal = data[col][data.Label == "s"].to_numpy()
    mask_signal = signal > -999
    background = data[col][data.Label == "b"].to_numpy()
    mask_background = background > -999
    xmin = min(np.min(background[mask_background]), np.min(signal[mask_signal]))
    xmax = max(np.max(background[mask_background]), np.max(signal[mask_signal]))

    ax.hist(signal[mask_signal], bins=np.linspace(xmin, xmax, 51), alpha=0.5,␣
 ↪label="signal", density=True)
    ax.hist(background[mask_background], bins=np.linspace(xmin, xmax, 51),␣
 ↪alpha=0.5, label="background", density=True)

    ax.set_xlabel(col, fontsize=40)
    ax.set_xlabel(col, fontsize=40)
    ax.tick_params(axis="both", which="major", labelsize=40)
    ax.legend(fontsize=40)
plt.tight_layout()
plt.show()
```

```
<Figure size 640x480 with 0 Axes>
```


signal
background
DER_mass_MMC
DER_mass_transverse_met_lep
DER_mass_vis
DER_pt_h
DER_deltaeta_jet_jet
DER_mass_jet_jet
DER_prodeta_jet_jet
DER_deltar_tau_lep
DER_pt_tot
DER_sum_pt
DER_pt_ratio_lep_tau
DER_met_phi_centrality
DER_lep_eta_centrality
PRI_tau_pt
PRI_tau_eta
PRI_tau_phi
PRI_lep_pt
PRI_lep_eta
PRI_lep_phi
PRI_met
PRI_met_phi
PRI_met_sumet
PRI_jet_num
PRI_jet_leading_pt
PRI_jet_leading_eta
PRI_jet_leading_phi
PRI_jet_subleading_pt
PRI_jet_subleading_eta
PRI_jet_subleading_phi
PRI_jet_all_pt

Format data: Now we should get the data into an XGBoost-friendly format. We can create DMatrix objects that will be used to train the BDT model. For now, we'll use all 30 of the features for training.

First, we'll slice up the data into training and testing sets. Here, we take 20% for the test set, which is arbitrary.

In this file, all samples are independent and ordered randomly, so we can just grab a chunk. Check out Scikit-learn Cross-validation for dividing up samples responsibly.

We can also change the data type of the `Label` column to the Pandas type `category` for easier use later.

[13]:
```
data["Label"] = data.Label.astype("category")
```

[14]:
```
data_train = data[:8000]
data_test = data[8000:]
```

Check to make sure we did it right:

[15]:
```
print(f"Number of training samples: {len(data_train)}")
print(f"Number of testing samples: {len(data_test)}")
print()
print(f"Number of signal events in training set: {len(data_train[data_train.
 ↪Label == 's'])}")
print(f"Number of background events in training set: {len(data_train[data_train.
 ↪Label == 'b'])}")
print(
    f"Fraction signal: {len(data_train[data_train.Label == 's'])/
 ↪(len(data_train[data_train.Label == 's']) + len(data_train[data_train.Label␣
 ↪== 'b']))}"
)
```

```
Number of training samples: 8000
Number of testing samples: 2000

Number of signal events in training set: 2688
Number of background events in training set: 5312
Fraction signal: 0.336
```

The `DMatrix` object takes as arguments: - `data`: the features - `label`: `1/0` or `True/False` for binary data (we have to convert our label to boolean from string `"s"/"b"`) - `missing`: how missing values are represented (here as `-999.0`) - `feature_names`: the names of all the features (optional)

[16]:
```
feature_names = list(data.columns[1:-2])  # we skip the first and last two␣
 ↪columns because they are the ID, weight, and label
```

```
print(len(feature_names))

train = xgb.DMatrix(
    data=data_train[feature_names], label=data_train.Label.cat.codes,␣
 ↪missing=-999.0, feature_names=feature_names
)
test = xgb.DMatrix(
    data=data_test[feature_names], label=data_test.Label.cat.codes,␣
 ↪missing=-999.0, feature_names=feature_names
)
```

```
30
```

Check if we did it right:

[17]:
```
print(f"Number of training samples: {train.num_row()}")
print(f"Number of testing samples: {test.num_row()}")
print()
print(f"Number of signal events in training set: {len(np.where(train.
 ↪get_label())[0])}")
```

```
Number of training samples: 8000
Number of testing samples: 2000

Number of signal events in training set: 2688
```

2.1.3 Make the model

Set hyperparameters: The XGBoost hyperparameters are defined here. For a nice description of what they all mean, and tips on tuning them, see this guide.

In general, the tunable parameters in XGBoost are the ones you would see in other gradient boosting libraries. Here, they fall into three categories: 1. General parameters: e.g., which booster to use, number of threads. We won't mess with these here. 2. Booster parameters: Tune the actual boosting, e.g., learning rate. These are the ones to optimize. 3. Learning task parameters: Define the objective function and the evaluation metrics.

Here, we will use the defaults for most parameters and just set a few to see how it's done. The parameters are passed in as a dictionary or list of pairs.

Make the parameter dictionary:

[18]:
```
param = {}

param["seed"] = 42  # set seed for reproducibility

# Booster parameters
param["eta"] = 0.1  # learning rate
param["max_depth"] = 10  # maximum depth of a tree
param["subsample"] = 0.8  # fraction of events to train tree on
```

```
param["colsample_bytree"] = 0.8  # fraction of features to train tree on

# Learning task parameters
param["objective"] = "binary:logistic"  # objective function
param["eval_metric"] = "error"  # evaluation metric for cross validation, note:␣
 ↪last one is used for early stopping
param = list(param.items())

num_trees = 100  # number of trees to make
```

First, we set the booster parameters. Again, we just chose a few here to experiment with. These are the parameters to tune to optimize your model. Generally, there is a trade off between speed and accuracy. 1. `eta` is the learning rate. It determines how much to change the data weights after each boosting iteration. The default is 0.3. 2. `max_depth` is the maximum depth of any tree. The default is 6. 3. `subsample` is the fraction of events used to train each new tree. These events are randomly sampled each iteration from the whole sample set. The default is 1 (use every event for each tree). 4. `colsample_bytree` is the fraction of features available to train each new tree. These features are randomly sampled each iteration from the whole feature set. The default is 1.

Next, we set the learning objective to `binary:logistic`. So, we have two classes that we want to score from 0 to 1. The `eval_metric` parameters set what we want to monitor when doing cross validation. (We aren't doing cross validation in this example, but we should be!) If you want to watch more than one metric, `param` must be a list of pairs, instead of a dict. Otherwise, we would just keep resetting the same parameter.

Last, we set the number of trees to 100. Usually, you would set this number high, and choose a cut off point based on the cross validation. The number of trees is the same as the number of iterations.

2.1.4 Now train!

[19]:
```
booster = xgb.train(param, train, num_boost_round=num_trees)
```

We now have a trained model. The next step is to look at it's performance and try to improve the model if we need to. We can try to improve it by improving/adding features, adding more training data, using more boosting iterations, or tuning the hyperparameters (ideally in that order).

First, let's look at how it does on the test set:

[20]:
```
print(booster.eval(test))
```

```
[0]     eval-error:0.17299999999999999
```

Okay, now we get the trained model, but how does it perform on the test set? These are the evaluation metrics that we stored in the parameter set.

It's pretty hard to interpret the performance of a classifier from a few number. So, let's look at the predictions for the entire test set.

3 Part 1

Use the trained model (booster) to make predictions on the test set.

[25]:
```
# Implement your code here
predictions = booster.predict(test) # (model's predictions on the test set)
labels = test.get_label().astype(int) # (the truth labels from the test set)
```

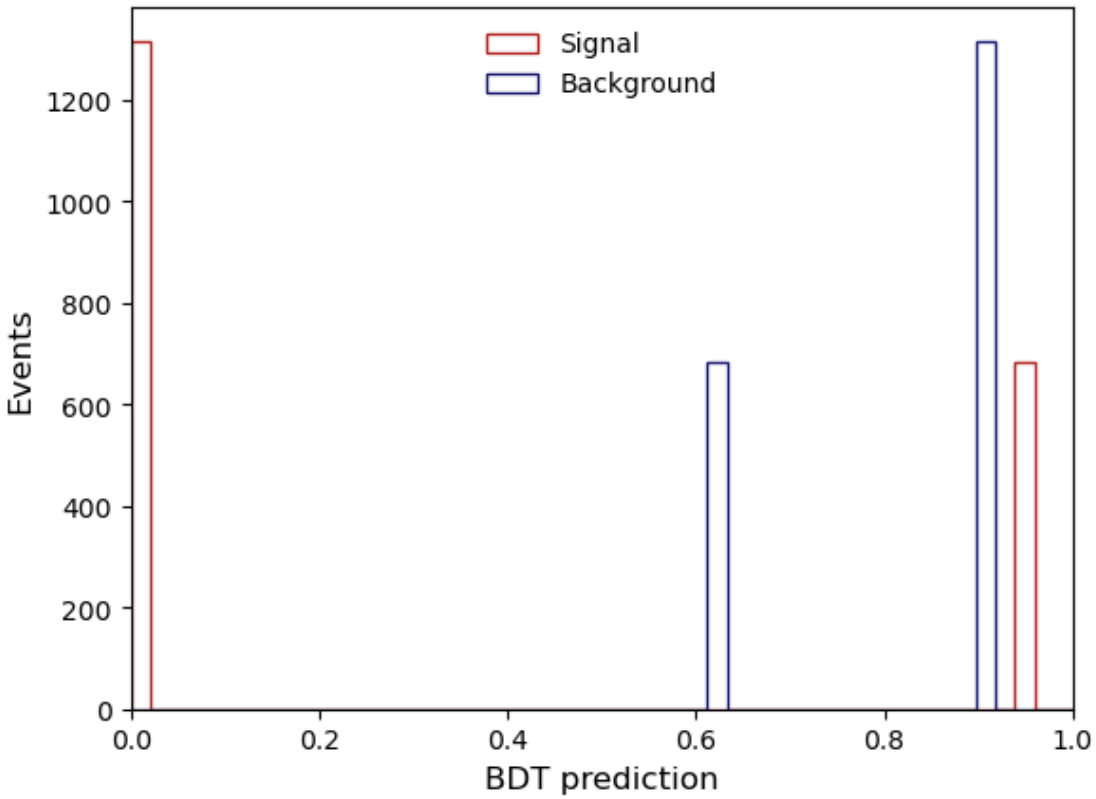
[28]:
```
labels
```

[28]:
```
array([0, 0, 1, …, 0, 0, 0])
```

[26]:
```
predictions
```

[26]:
```
array([0.00757499, 0.9505924 , 0.31410593, …, 0.19513501, 0.6230965 ,
       0.9080286 ], dtype=float32)
```

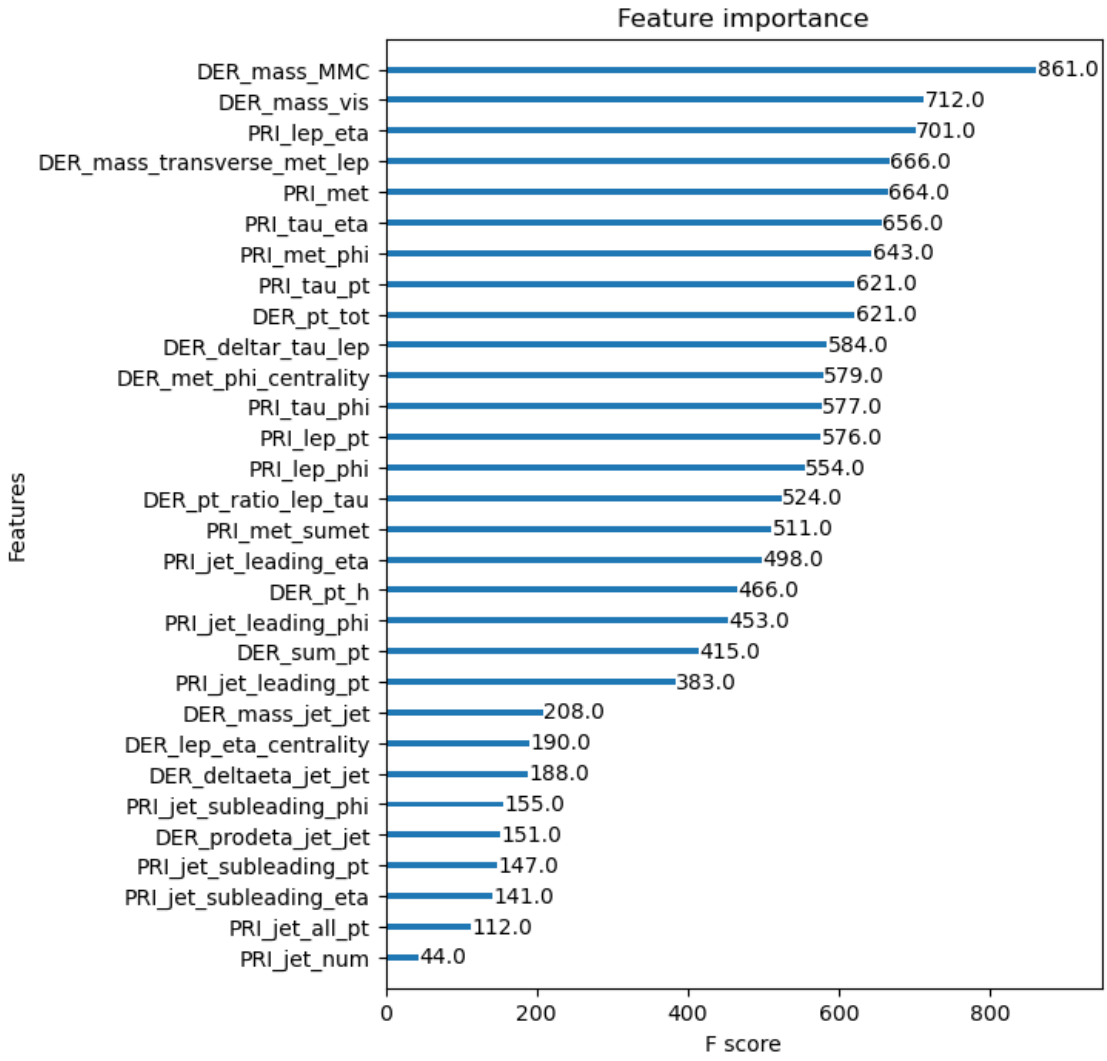
[31]:
```
# Plot signal and background predictions, separately
plt.figure()
plt.hist(predictions[labels], bins=np.linspace(0, 1, 50), histtype="step",␣
 ↪color="firebrick", label="Signal")
plt.hist(predictions[~labels], bins=np.linspace(0, 1, 50), histtype="step",␣
 ↪color="midnightblue", label="Background")
# Make the plot readable
plt.xlabel("BDT prediction", fontsize=12)
plt.ylabel("Events", fontsize=12)
plt.legend(frameon=False)
plt.xlim(0, 1)
plt.show()
```



It's also very informative to look at the importance of each feature. The "F score" is the number of times each feature is used to split the data over all the trees (times the weight of that tree).

There is a built-in function in the XGBoost Python API to easily plot this.

[29]:
```
fig, ax = plt.subplots(figsize=(6, 8))
xgb.plot_importance(booster, ax=ax, grid=False)
plt.show()
```

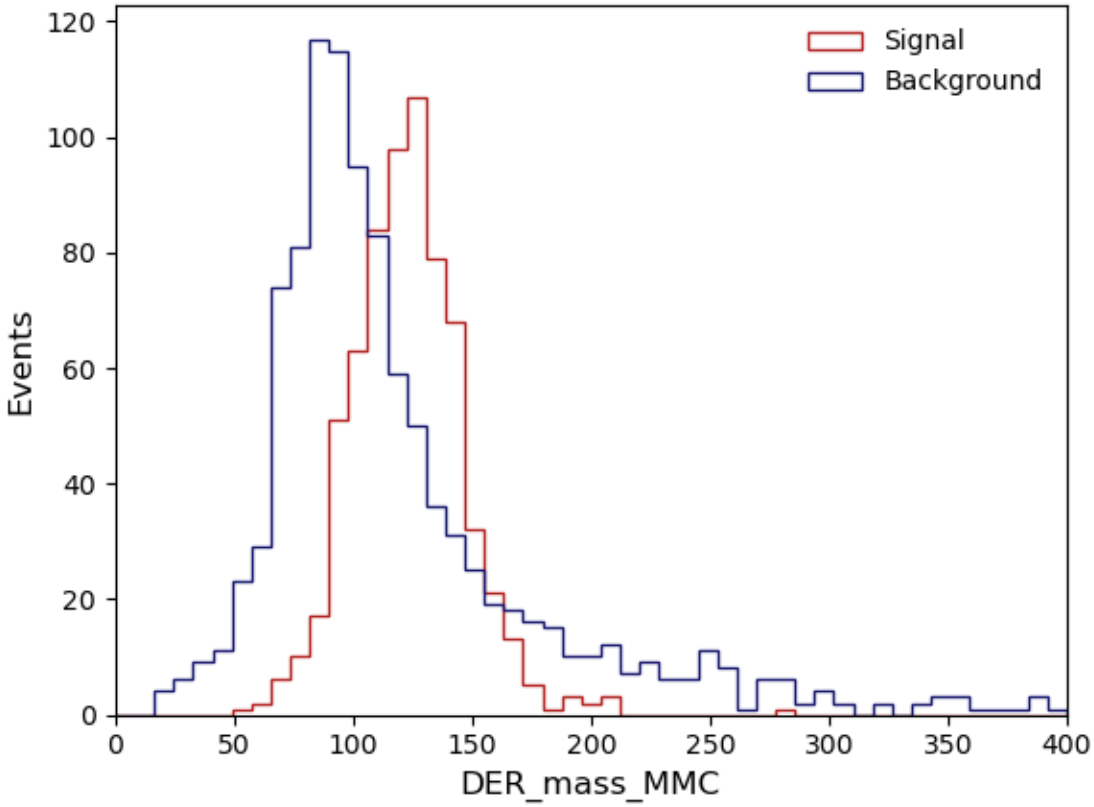


The feature that was used the most was `DER_mass_MMC`.

We can plot how this feature is distributed for the signal and background.

[32]:
```
plt.figure()
plt.hist(
    data_test.DER_mass_MMC[data_test.Label == "s"],
    bins=np.linspace(0, 400, 50),
    histtype="step",
    color="firebrick",
    label="Signal",
)
plt.hist(
    data_test.DER_mass_MMC[data_test.Label == "b"],
    bins=np.linspace(0, 400, 50),
```

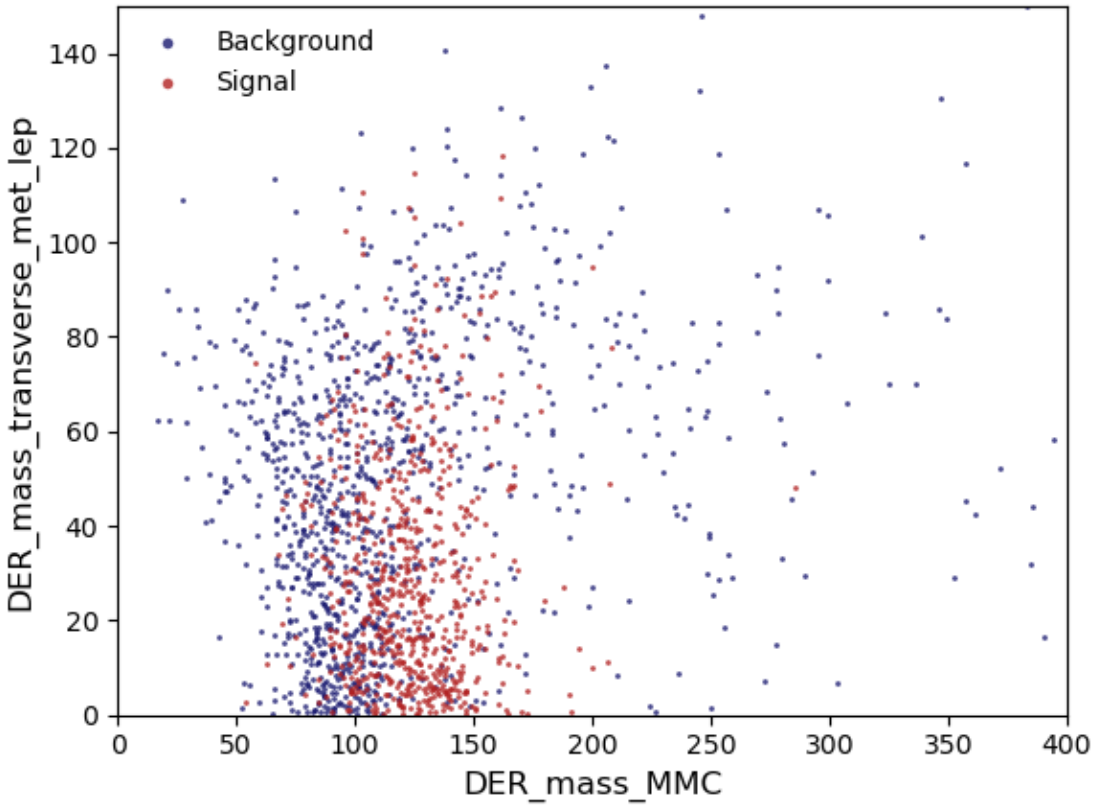
```
    histtype="step",
    color="midnightblue",
    label="Background",
)

plt.xlim(0, 400)
plt.xlabel("DER_mass_MMC", fontsize=12)
plt.ylabel("Events", fontsize=12)
plt.legend(frameon=False)
plt.show()
```



This variable is physically significant because it represents an estimate of the Higgs boson mass. For signal, it is expected to peak at 125 GeV. We can also plot it with one of the next most important features `DER_mass_transverse_met_lep`. Note: the exact ranking of features can depend on the random seed and other hyperparameters.

[33]:
```
plt.figure()

mask_b = np.array(data_test.Label == "b")
mask_s = np.array(data_test.Label == "s")
```

```
DER_mass_MMC = np.array(data_test.DER_mass_MMC)
DER_mass_transverse_met_lep = np.array(data_test.DER_mass_transverse_met_lep)

plt.plot(
    DER_mass_MMC[mask_b],
    DER_mass_transverse_met_lep[mask_b],
    "o",
    markersize=2,
    color="midnightblue",
    markeredgewidth=0,
    alpha=0.8,
    label="Background",
)
plt.plot(
    DER_mass_MMC[mask_s],
    DER_mass_transverse_met_lep[mask_s],
    "o",
    markersize=2,
    color="firebrick",
    markeredgewidth=0,
    alpha=0.8,
    label="Signal",
)
plt.xlim(0, 400)
plt.ylim(0, 150)
plt.xlabel("DER_mass_MMC", fontsize=12)
plt.ylabel("DER_mass_transverse_met_lep", fontsize=12)
plt.legend(frameon=False, numpoints=1, markerscale=2)
plt.show()
```



3.1 Part 2

Calculate the precision, recall and F1-score as we learned from the lecture. Assuming using the default threshold 0.5.

1. Precision measures how many of the positive predictions made by the model are actually correct.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

2. Recall (also called sensitivity or true positive rate) measures how many of the actual positive cases were correctly identified by the model.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

3. F1-score is the harmonic mean of precision and recall, providing a single metric to balance the trade-off between them. It ranges from 0 to 1, where 1 indicates perfect precision and recall.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

[37]:
```
# Implement your code below
from sklearn.metrics import confusion_matrix
```

```
threshold = 0.5
y_pred = (predictions >= threshold).astype(int)

tn, fp, fn, tp = confusion_matrix(labels, y_pred).ravel()
print(f"True Positives (TP):  {tp}")
print(f"False Positives (FP): {fp}")
print(f"True Negatives (TN):  {tn}")
print(f"False Negatives (FN): {fn}")

precision = tp / (tp + fp)
recall = tp / (tp + fn)
f1_score = 2 * precision * recall / (precision + recall)
```

```
True Positives (TP):  475
False Positives (FP): 137
True Negatives (TN):  1179
False Negatives (FN): 209
```

[38]:
```
# Check results
import math
assert math.isclose(precision, 0.7761, rel_tol=1e-4), f"Precision not correct!"
assert math.isclose(recall, 0.6944, rel_tol=1e-4), f"Recall not correct!"
assert math.isclose(f1_score, 0.7330, rel_tol=1e-4), f"F1-Score not correct!"
```

3.2 Part 3

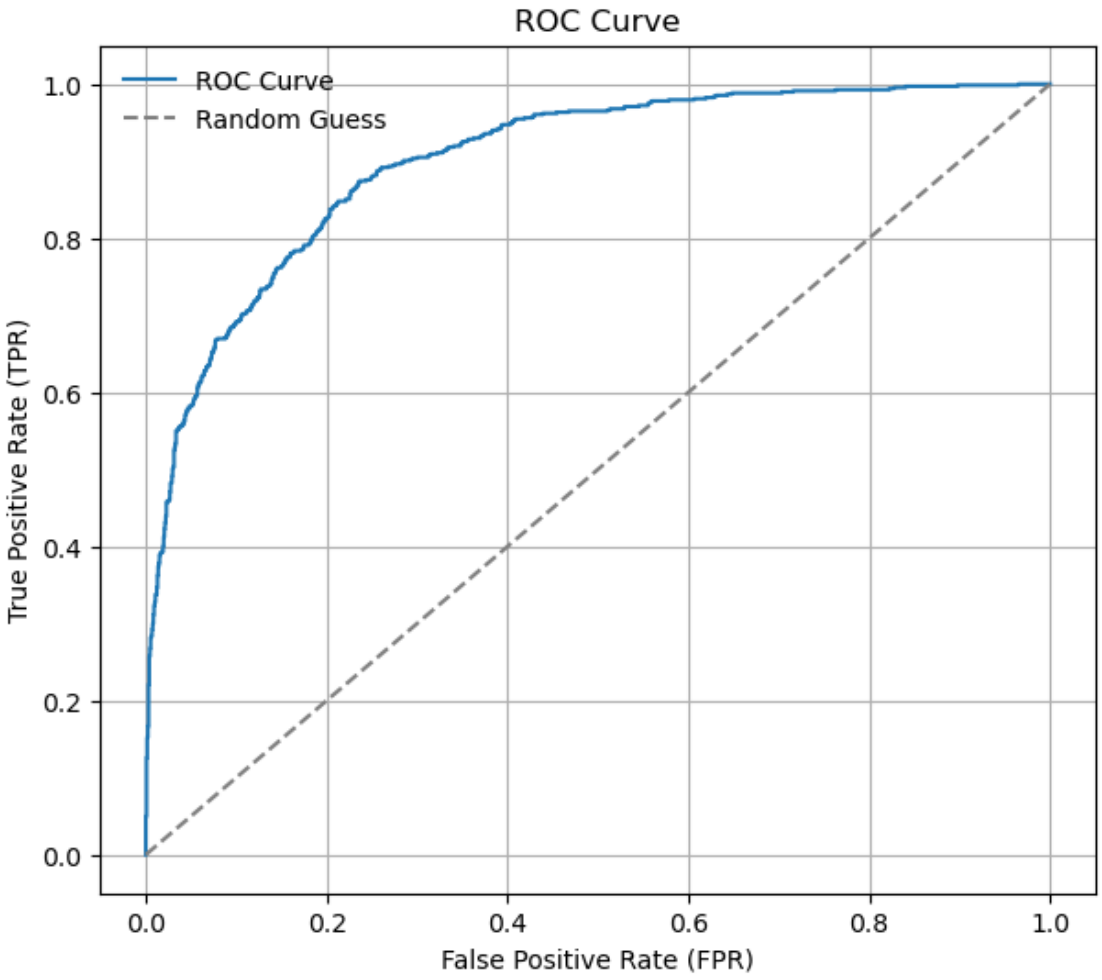
Plot the ROC curve as we discussed during class with your own calculations, do not use the roc_curve library

[48]:
```
import numpy as np
import matplotlib.pyplot as plt

def compute_roc_curve(predictions, labels):
    # Input: predictions, labels
    # Plot the ROC curve as we discussed during class with your own␣
 ↪calculations, do not use the roc_curve library
    # Output: fpr, tpr: arrays of FPRs and TPRs

    # Implement your code below
    predictions = np.array(predictions)
    labels = np.array(labels)


    # Sort predictions and corresponding labels in descending order
    sorted_indices = np.argsort(-predictions)
    sorted_predictions = predictions[sorted_indices]
    sorted_labels = labels[sorted_indices]
```

```
    unique_thresholds = np.unique(sorted_predictions)[::-1]

    tpr_list = []
    fpr_list = []

    P = np.sum(labels)
    N = len(labels) - P

    TP, FP = 0, 0
    FN, TN = P, N

    for threshold in unique_thresholds:
        # Find all instances with prediction >= threshold
        while len(sorted_predictions) > 0 and sorted_predictions[0] >=␣
↪threshold:
            current_label = sorted_labels[0]
            if current_label == 1:
                TP += 1
                FN -= 1
            else:
                FP += 1
                TN -= 1
            # Remove the processed instance
            sorted_predictions = sorted_predictions[1:]
            sorted_labels = sorted_labels[1:]

        # Calculate TPR and FPR
        TPR = TP / P if P != 0 else 0
        FPR = FP / N if N != 0 else 0

        tpr_list.append(TPR)
        fpr_list.append(FPR)

    # Append (1,1) to complete the ROC curve
    tpr_list = [0.0] + tpr_list + [1.0]
    fpr_list = [0.0] + fpr_list + [1.0]

    return np.array(fpr_list), np.array(tpr_list)
```

[49]:
```
fpr, tpr = compute_roc_curve(predictions, labels)

# Plot ROC curve
plt.figure(figsize=(7, 6))
plt.plot(fpr, tpr, linestyle='-', label="ROC Curve")
plt.plot([0, 1], [0, 1], linestyle='--', color='gray', label="Random Guess")
plt.xlabel("False Positive Rate (FPR)")
```

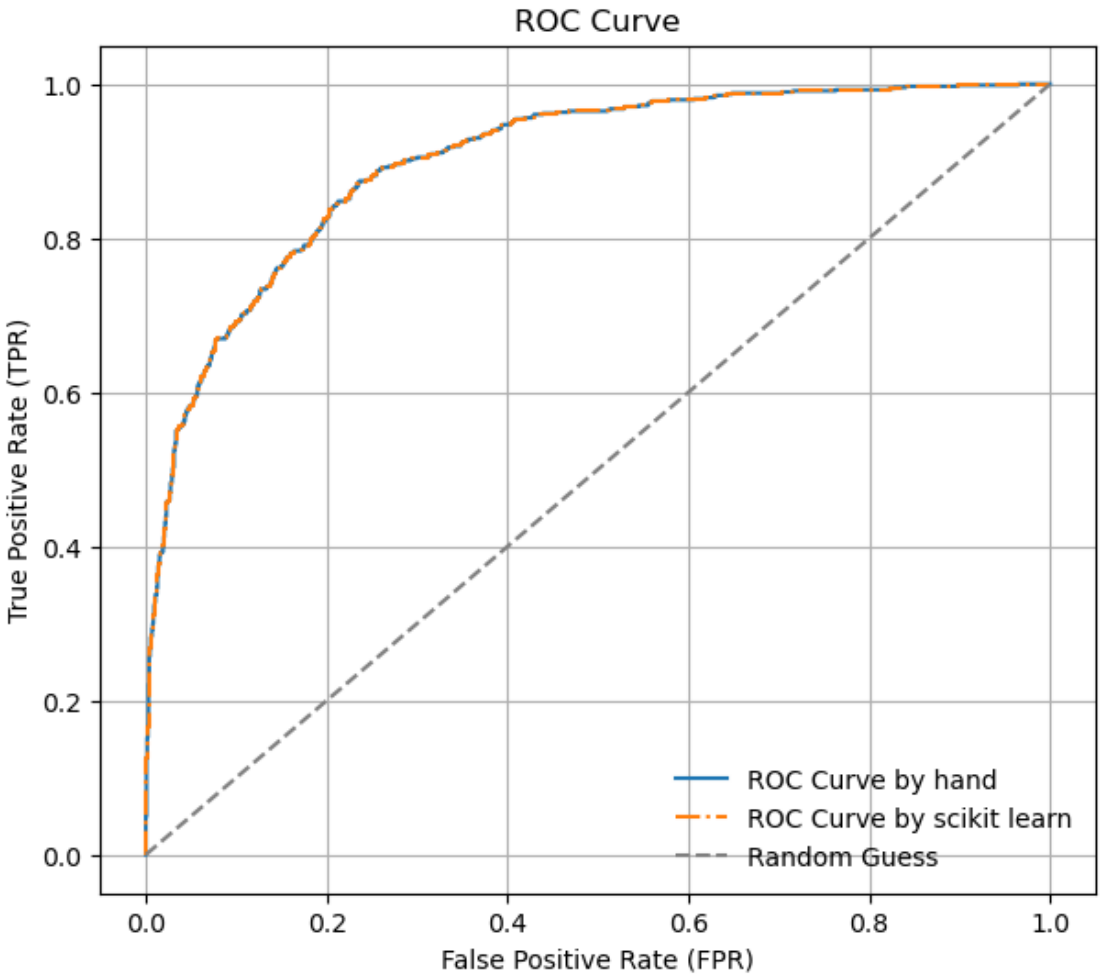
```
plt.ylabel("True Positive Rate (TPR)")
plt.title("ROC Curve")
plt.legend(frameon=False)
plt.grid()
plt.show()
```



3.3 Validation: plot using scikit learn

[51]:
```
from sklearn.metrics import confusion_matrix, roc_curve, roc_auc_score
fpr_sc, tpr_sc, roc_thresholds = roc_curve(labels, predictions)

# Plot ROC curve
plt.figure(figsize=(7, 6))
plt.plot(fpr, tpr, linestyle='-', label="ROC Curve by hand")
plt.plot(fpr_sc, tpr_sc, linestyle='-.', label="ROC Curve by scikit learn")
plt.plot([0, 1], [0, 1], linestyle='--', color='gray', label="Random Guess")
```

```
plt.xlabel("False Positive Rate (FPR)")
plt.ylabel("True Positive Rate (TPR)")
plt.title("ROC Curve")
plt.legend(frameon=False)
plt.grid()
plt.show()
```



[ ]: