In [3]:
```
import warnings
warnings.filterwarnings('ignore')
```

Dataset Sources

Beat Acute Myeloid Leukemia (AML) 1.0 was accessed on 13Mar2023 from https://registry.opendata.aws/beataml. OHSU BeatAML Datasets Link: https://ctd2-data.nci.nih.gov/Public/OHSU-1/BeatAML_Waves1_2/

OpenCell Datasets Link: https://opencell.czbiohub.org/download

Check Pre-Requisites from the `01_setup/` Folder

In [4]:
```
%store -r setup_instance_check_passed
```

In [5]:
```
try:
    setup_instance_check_passed
except NameError:
    print("+++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Instance Check.")
    print("+++++++++++++++++++++++++++++++")
```

In [6]:
```
print(setup_instance_check_passed)
```

```
True
```

In [7]:
```
%store -r setup_dependencies_passed
```

In [8]:
```
try:
    setup_dependencies_passed
except NameError:
    print("+++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Setup Dependencies.")
    print("+++++++++++++++++++++++++++++++")
```

In [9]:
```
print(setup_dependencies_passed)
```

```
True
```

In [10]:
```
%store -r setup_s3_bucket_passed
```

In [11]:
```
try:
    setup_s3_bucket_passed
except NameError:
    print("+++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Setup S3 Bucket.")
    print("+++++++++++++++++++++++++++++++")
```

In [12]:
```
print(setup_s3_bucket_passed)
```

```
True
```

In [13]:
```
%store -r setup_iam_roles_passed
```

In [14]:
```
try:
    setup_iam_roles_passed
except NameError:
    print("+++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Setup IAM Roles.")
    print("+++++++++++++++++++++++++++++++")
```

In [15]:
```
print(setup_iam_roles_passed)
```

```
True
```

In [16]:
```
if not setup_instance_check_passed:
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Instance Check.")
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
if not setup_dependencies_passed:
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Setup Dependencies.")
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
if not setup_s3_bucket_passed:
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Setup S3 Bucket.")
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
if not setup_iam_roles_passed:
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Setup IAM Roles.")
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

In [17]:

```
import boto3
import sagemaker
import pandas as pd

sess = sagemaker.Session()
bucket = sess.default_bucket()
role = sagemaker.get_execution_role()
region = boto3.Session().region_name
account_id = boto3.client("sts").get_caller_identity().get("Account")

sm = boto3.Session().client(service_name="sagemaker", region_name=region)
```

S3 Original Dataset Location

Importing Raw Datasets from AWS S3. Use the AWS Command Line Interface (CLI) to list the S3 bucket content using the following CLI commands:

In [18]:

```
!aws s3 ls s3://team4rawdatasets/CSV/Input/OHSU_BeatAML_ClinicalSummary/
```

```
2023-03-21 02:37:21          0
2023-03-27 02:40:07     714614 OHSU_BeatAMLWaves1_2_Tyner_ClinicalSummary.txt
```

In [19]:

```
!aws s3 ls s3://team4rawdatasets/CSV/Input/OpenCell_ProteinInteraction/
```

```
2023-03-21 02:37:38          0
2023-03-21 02:38:40    4568928 opencell-protein-interactions.csv
```

Set S3 Source Location

In [20]:

```
#BeatAML Clinical Summary
s3_public_path_clsm = "s3://team4rawdatasets/CSV/Input/OHSU_BeatAML_ClinicalSummary/"
```

In [21]:

```
%store s3_public_path_clsm
```

```
Stored 's3_public_path_clsm' (str)
```

In [22]:

```
print(s3_public_path_clsm)
```

```
s3://team4rawdatasets/CSV/Input/OHSU_BeatAML_ClinicalSummary/
```

In [23]:
```
!aws s3 ls $s3_public_path_clsm
```

```
2023-03-21 02:37:21          0 
2023-03-27 02:40:07     714614 OHSU_BeatAMLWaves1_2_Tyner_ClinicalSummary.txt
```

In [24]:
```
#BeatAML OpenCell Protein Interaction
s3_public_path_pi = "s3://team4rawdatasets/CSV/Input/OpenCell_ProteinInteraction/"
```

In [25]:
```
%store s3_public_path_pi
```

```
Stored 's3_public_path_pi' (str)
```

In [26]:
```
print(s3_public_path_pi)
```

```
s3://team4rawdatasets/CSV/Input/OpenCell_ProteinInteraction/
```

In [27]:
```
!aws s3 ls $s3_public_path_pi
```

```
2023-03-21 02:37:38          0 
2023-03-21 02:38:40    4568928 opencell-protein-interactions.csv
```

In [28]:
```
from IPython.core.display import display, HTML

display(
    HTML(
        '<b>Review <a target="blank" href="https://s3.console.aws.amazon.com/s3/buckets/team4rawdatasets?prefix=CSV/I
            region, account_id, region
        )
    )
)
```

Review S3 Bucket

Athena

Athena Database

PyAthena is a Python DB API 2.0 (PEP 249) compliant client for Amazon Athena.

In [29]:

```
!pip install --disable-pip-version-check -q PyAthena==2.1.0
from pyathena import connect
```

```
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system p
ackage manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
```

In [30]:

```
ingest_create_athena_db_passed = False
```

In [31]:

```
database_name = "bcr"
```

In [32]:

```
# Set S3 staging directory -- this is a temporary directory used for Athena queries
s3_staging_dir = "s3://{0}/athena/staging".format(bucket)
```

In [33]:

```
conn = connect(region_name=region, s3_staging_dir=s3_staging_dir)
```

In [34]:

```
statement0 = "CREATE DATABASE IF NOT EXISTS {}".format(database_name)
print(statement0)
```

```
CREATE DATABASE IF NOT EXISTS bcr
```

In [35]:

```
pd.read_sql(statement0, conn)
```

Out[35]:

Verify The Database Has Been Created Succesfully

In [36]:

```
statement00 = "SHOW DATABASES"

df_show = pd.read_sql(statement00, conn)
df_show.head(5)
```

Out[36]:

	database_name
0	bcr
1	default
2	dsoaws
3	sagemaker_featurestore

In [37]:
```python
if database_name in df_show.values:
    ingest_create_athena_db_passed = True
```

In [38]:
```python
%store ingest_create_athena_db_passed
```

```
Stored 'ingest_create_athena_db_passed' (bool)
```

Athena Table Created Through AWS Glue Crawler

In [39]:
```python
from IPython.core.display import display, HTML

display(
    HTML(
        '<b>Review <a target="top" href="https://us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data
            region
        )
    )
)
```

Review AWS Glue Catalog

Athena Sample Query

In [40]:
```python
# Set Athena database & table
table_clsm = "ohsu_beataml_clinicalsummary"
table_pi = "opencell_proteininteraction"
```

In [41]:
```python
#Athena SQL Code
statement1 = """
SELECT *
FROM {}.{}
""".format(
    database_name, table_pi
)

print(statement1)
```

```
SELECT *
FROM bcr.opencell_proteininteraction
```

In [42]:

```
pi = pd.read_sql(statement1, conn)
pi.head(5)
```

Out[42]:

	target_gene_name	interactor_gene_name	target_ensg_id	interactor_ensg_id	interactor_uniprot_ids
0	CAPZB	LIN7C	ENSG00000077549	ENSG00000148943	Q9NUP9;G3V1D4
1	CAPZB	LMO7	ENSG00000077549	ENSG00000136153	Q8WWI1-3;Q8WWI1;Q8WWI1-2;Q8WWI1-4;J3KP06;F8WD2...
2	CAPZB	LONP1	ENSG00000077549	ENSG00000196365	K7EJE8;K7EKE6;P36776-3;P36776-2;P36776;K7ER27
3	CAPZB	LRCH2	ENSG00000077549	ENSG00000130224	Q5VUJ6-2;Q5VUJ6
4	CAPZB	LRPPRC	ENSG00000077549	ENSG00000138095	P42704;C9JCA9;B8ZZ38;A0A0C4DG06;H7C3W8

In [43]:

```
if not pi.empty:
    print("[OK]")
else:
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOUR DATA HAS NOT BEEN CONVERTED TO PARQUET. LOOK IN PREVIOUS CELLS TO FIND THE ISSUE.")
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

```
[OK]
```

Data cleaning

Import Tools:

In [44]:

```
!pip install klib
```

```
Requirement already satisfied: klib in /opt/conda/lib/python3.7/site-packages (1.0.1)
Requirement already satisfied: pandas<2.0.0,>=1.1.2 in /opt/conda/lib/python3.7/site-packages (from klib) (1.3.5)
Requirement already satisfied: scipy<2.0.0,>=1.1.0 in /opt/conda/lib/python3.7/site-packages (from klib) (1.4.1)
Requirement already satisfied: numpy<2.0.0,>=1.16.3 in /opt/conda/lib/python3.7/site-packages (from klib) (1.21.6)
Requirement already satisfied: seaborn<0.12.0,>=0.11.1 in /opt/conda/lib/python3.7/site-packages (from klib) (0.11.
2)
Requirement already satisfied: Jinja2<4.0.0,>=3.0.3 in /opt/conda/lib/python3.7/site-packages (from klib) (3.1.2)
Requirement already satisfied: matplotlib<4.0.0,>=3.0.3 in /opt/conda/lib/python3.7/site-packages (from klib) (3.1.
3)
Requirement already satisfied: MarkupSafe>=2.0 in /opt/conda/lib/python3.7/site-packages (from Jinja2<4.0.0,>=3.0.3-
>klib) (2.1.2)
Requirement already satisfied: python-dateutil>=2.1 in /opt/conda/lib/python3.7/site-packages (from matplotlib<4.0.
0,>=3.0.3->klib) (2.8.2)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/conda/lib/python3.7/site-packages (from matplotlib<4.0.0,>=
3.0.3->klib) (1.1.0)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.7/site-packages (from matplotlib<4.0.0,>=3.0.3
->klib) (0.10.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /opt/conda/lib/python3.7/site-packages (f
rom matplotlib<4.0.0,>=3.0.3->klib) (2.4.6)
Requirement already satisfied: pytz>=2017.3 in /opt/conda/lib/python3.7/site-packages (from pandas<2.0.0,>=1.1.2->kl
ib) (2019.3)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages (from cycler>=0.10->matplotlib<4.0.0,>=
3.0.3->klib) (1.14.0)
Requirement already satisfied: setuptools in /opt/conda/lib/python3.7/site-packages (from kiwisolver>=1.0.1->matplot
lib<4.0.0,>=3.0.3->klib) (59.3.0)
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system p
ackage manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

[notice] A new release of pip is available: 23.0.1 -> 23.1
[notice] To update, run: pip install --upgrade pip
```

In [45]:

```
import numpy as np
import seaborn as sns
import klib
import matplotlib.pyplot as plt

%matplotlib inline
%config InlineBackend.figure_format='retina'
```

BeatAML Clinical Summary

OHSU BeatAML Clinical Summary Table

In [46]:

```
# SQL statement
statement2 = """
SELECT *
FROM {}.{}
""".format(
    database_name, table_clsm
)

print(statement2)
```

```
SELECT *
FROM bcr.ohsu_beataml_clinicalsummary
```

In [47]:

```
clsm = pd.read_sql(statement2, conn)
clsm.head(5)
```

Out[47]:

	labid	patientid	consensus_sex	inferred_sex	inferred_ethnicity	centerid	cebpa_biallelic	ageatdiagnosis	isrelapse	isdenovo	...	st
0	09-00705	163	Male	Male	White	1	n	73.0	False	True	...	
1	10-00136	174	Male	Male	White	1	n	69.0	False	True	...	
2	10-00172	175	Female	Male	White	1	n	59.0	False	True	...	
3	10-00507	45	Female	Female	White	1	n	70.0	False	True	...	
4	10-00542	174	Male	Male	White	1	n	69.0	True	False	...	

5 rows × 159 columns

In [48]:

```
clsm = clsm.replace('', np.NAN)
clsm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Columns: 159 entries, labid to zrsr2
dtypes: bool(9), float64(22), int64(7), object(121)
memory usage: 793.5+ KB
```

In [49]:

```
clsm.info(2)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 159 columns):
 #    Column                                          Dtype  
---   ------                                          -----  
 0    labid                                           object 
 1    patientid                                       int64  
 2    consensus_sex                                   object 
 3    inferred_sex                                    object 
 4    inferred_ethnicity                              object 
 5    centerid                                        int64  
 6    cebpa_biallelic                                 object 
 7    ageatdiagnosis                                  float64
 8    isrelapse                                       bool   
 9    isdenovo                                        bool   
 10   istransformed                                   bool   
 11   finalfusion                                     object 
 12   specificdxatacquisition_mdsmpn                  bool   
 13   nonaml_mdsmpn_specificdxatacquisition           bool   
 14   priormalignancynonmyeloid                       object 
 15   priormalignancytype                             object 
 16   cumulativechemo                                 object 
 17   priormalignancyradiationtx                      object 
 18   priormds                                        object 
 19   priormdsmorethantwomths                         object 
 20   priormdsmpn                                     object 
 21   priormdsmpnmorethantwomths                      object 
 22   priormpn                                        object 
 23   priormpnmorethantwomths                         object 
 24   dxatinclusion                                   object 
 25   specificdxatinclusion                           object 
 26   eln2017                                         object 
 27   eln2008                                         object 
 28   dxatspecimenacquisition                         object 
 29   specificdxatacquisition                         object 
 30   ageatspecimenacquisition                        float64
 31   timeofsamplecollectionrelativetoinclusion       int64  
 32   specimengroups                                  object 
 33   specimentype                                    object 
 34   rnaseq                                          object 
 35   exomeseq                                        object 
 36   totaldrug                                       object 
 37   rnaseqanalysis                                  object 
 38   analysisexomeseq                                object 
```

```
39   analysisdrug                                   object
40   cumulativetreatmenttypecount                   int64
41   cumulativetreatmenttypes                       object
42   cumulativetreatmentregimencount                int64
43   cumulativetreatmentregimens                    object
44   cumulativetreatmentstagecount                  int64
45   cumulativetreatmentstages                      object
46   responsetoinductiontx                          object
47   typeinductiontx                                object
48   responsedurationtoinductiontx                  float64
49   mostrecenttreatmenttype                        object
50   currentregimen                                 object
51   currentstage                                   object
52   mostrecenttreatmentduration                    float64
53   vitalstatus                                    object
54   overallsurvival                                float64
55   causeofdeath                                   object
56   any_different_labs                             bool
57   any_different_labs_also_beataml                bool
58   different_lab_ids                              object
59   different_id_karyotype_interval                int64
60   %.basophils.in.pb                              float64
61   %.blasts.in.bm                                 object
62   %.blasts.in.pb                                 object
63   %.eosinophils.in.pb                            float64
64   %.immature.granulocytes.in.pb                  float64
65   %.lymphocytes.in.pb                            float64
66   %.monocytes.in.pb                              float64
67   %.neutrophils.in.pb                            float64
68   %.nucleated.rbcs.in.pb                         float64
69   alt                                            object
70   ast                                            float64
71   albumin                                        float64
72   creatinine                                     float64
73   fab/blast.morphology                           object
74   hematocrit                                     float64
75   hemoglobin                                     float64
76   karyotype                                      object
77   ldh                                            float64
78   mcv                                            float64
79   other.cytogenetics                             object
80   platelet.count                                 float64
81   surface.antigens.(immunohistochemical.stains)  object
82   total.protein                                  float64
```

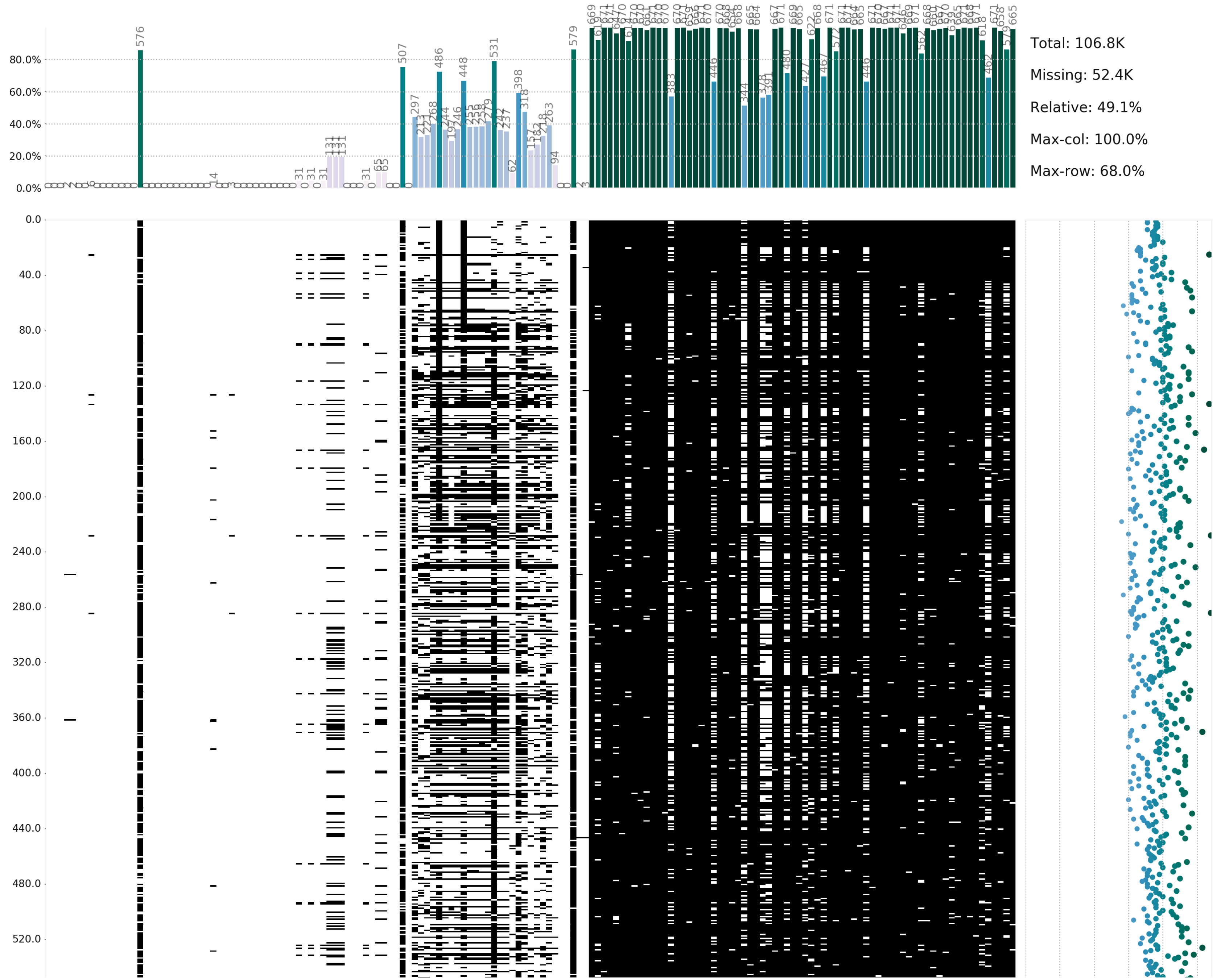
```
83   wbc.count                                      float64
84   any_different_cgs                              bool
85   any_different_cgs_also_beataml                 bool
86   different_cgs_lab_ids                          object
87   flt3-itd                                       object
88   npm1                                           object
89   abl1                                           object
90   asxl1                                          object
91   asxl2                                          object
92   atm                                            object
93   bcor                                           object
94   bcorl1                                         object
95   braf                                           object
96   brca2                                          object
97   calr                                           object
98   cbl                                            object
99   ccnd2                                          object
100  ccnd3                                          object
101  cd36                                           object
102  cebpa                                          object
103  chek2                                          object
104  ciita                                          object
105  crebbp                                         object
106  csf3r                                          object
107  ctcf                                           object
108  cux1                                           object
109  dnmt3a                                         object
110  ep300                                          object
111  etv6                                           object
112  ezh2                                           object
113  fbxw7                                          object
114  flt3                                           object
115  gata1                                          object
116  gata2                                          object
117  idh1                                           object
118  idh2                                           object
119  ikzf1                                          object
120  jak1                                           object
121  jak2                                           object
122  jak3                                           object
123  kdm6a                                          object
124  kit                                            object
125  kmt2a                                          object
126  kmt2d                                          object
```

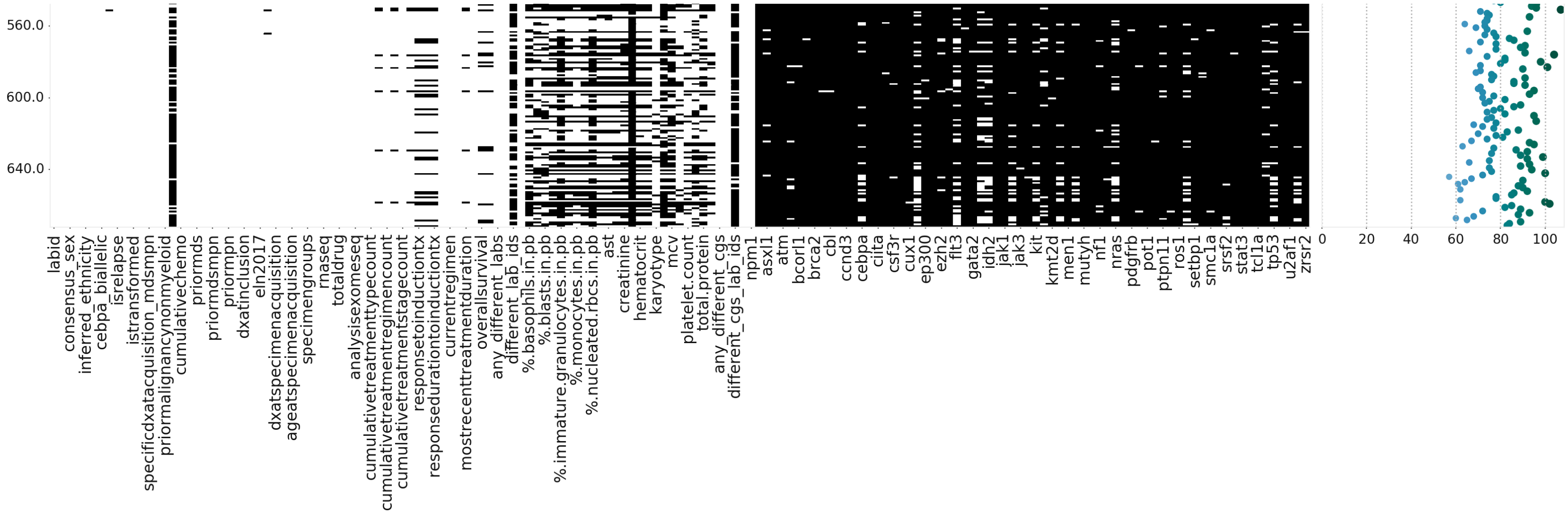
```
 127  kras                                          object
 128  men1                                          object
 129  mpl                                           object
 130  mutyh                                         object
 131  myd88                                         object
 132  nf1                                           object
 133  notch1                                        object
 134  nras                                          object
 135  pax5                                          object
 136  pdgfrb                                        object
 137  phf6                                          object
 138  pot1                                          object
 139  prdm1                                         object
 140  ptpn11                                        object
 141  rad21                                         object
 142  ros1                                          object
 143  runx1                                         object
 144  setbp1                                        object
 145  sf3b1                                         object
 146  smc1a                                         object
 147  socs1                                         object
 148  srsf2                                         object
 149  stag2                                         object
 150  stat3                                         object
 151  suz12                                         object
 152  tcl1a                                         object
 153  tet2                                          object
 154  tp53                                          object
 155  tyk2                                          object
 156  u2af1                                         object
 157  wt1                                           object
 158  zrsr2                                         object
dtypes: bool(9), float64(22), int64(7), object(121)
memory usage: 793.5+ KB
```

```
In [50]: klib.missingval_plot(clsm)
```

```
Out[50]: GridSpec(6, 6)
```

Missing value plot





Select Relevant Features

In [51]:

```
clsm_cut = pd.DataFrame(clsm[['labid', 'patientid', 'consensus_sex', 'inferred_ethnicity', 'isrelapse',
                              'istransformed', 'priormalignancynonmyeloid', 'priormds', 'priormdsmpn', 'priormpn',
                              'eln2017', 'dxatspecimenacquisition', 'vitalstatus', 'overallsurvival', '%.blasts.in.bm
                               '%.blasts.in.pb','flt3-itd', 'npm1']])

clsm_cut
```

```
Out[51]:
         labid  patientid  consensus_sex  inferred_ethnicity  isrelapse  istransformed  priormalignancynonmyeloid  priormds  priormdsmpn

  0   09-00705        163           Male               White      False          False                          n         n            n

  1   10-00136        174           Male               White      False          False                          n         n            n

  2   10-00172        175         Female               White      False          False                          n         n            n

  3   10-00507         45         Female               White      False          False                          n         n            n

  4   10-00542        174           Male               White       True          False                          n         n            n
...        ...        ...            ...                 ...        ...            ...                        ...       ...          ...
667   17-00072       4366           Male               White      False           True                          n         n            n

668   17-00077       4317         Female               White      False          False                          n         n            n

669   17-00093       4379         Female               Black      False           True                          n         n            n

670   17-00094       4380           Male               White      False           True                          n         n            n

671   17-00096       2747           Male               White      False           True                          n         n            y

672 rows × 18 columns
```

In [52]:

```
clsm_cut.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 18 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   labid                       672 non-null    object 
 1   patientid                   672 non-null    int64  
 2   consensus_sex               672 non-null    object 
 3   inferred_ethnicity          670 non-null    object 
 4   isrelapse                   672 non-null    bool   
 5   istransformed               672 non-null    bool   
 6   priormalignancynonmyeloid   672 non-null    object 
 7   priormds                    672 non-null    object 
 8   priormdsmpn                 672 non-null    object 
 9   priormpn                    672 non-null    object 
 10  eln2017                     672 non-null    object 
 11  dxatspecimenacquisition     672 non-null    object 
 12  vitalstatus                 672 non-null    object 
 13  overallsurvival             607 non-null    float64
 14  %.blasts.in.bm              459 non-null    object 
 15  %.blasts.in.pb              451 non-null    object 
 16  flt3-itd                    670 non-null    object 
 17  npm1                        669 non-null    object 
dtypes: bool(2), float64(1), int64(1), object(14)
memory usage: 85.4+ KB
```

In [53]:

```
clsm_cut.describe()
```

Out[53]:

	patientid	overallsurvival
count	672.000000	607.000000
mean	2088.020833	441.881384
std	973.372734	479.180429
min	17.000000	-1.000000
25%	1450.750000	167.000000
50%	2016.000000	323.000000
75%	2501.500000	555.000000
max	4380.000000	5305.000000

Attribute Tranformation

% Blasts Attributes Numerical Prep

%.blasts.in.bm Attribute:

In [54]:

```
clsm_cut['%.blasts.in.bm'].unique()
```

Out[54]:

```
array(['94', '80', '91', '97', '87', nan, '40', '75', '83', '95', '85',
       '90', '70', '92', '72', '68', '88', '36', '81', '93', '34', '77.5',
       '46', '65', '50', '76', '71', '60', '73', '55', '0.5', '30', '62',
       '18', '82', '28', '41', '64', '84', '21', '51', '17', '49.4', '32',
       '29', '25', '59.3', '66', '20', '52', '54', '22', '10', '12',
       '46.0', '13', '67', '39', '25.9', '45', '37', '78', '8', '3',
       '54.8', '74', '96', '4', '86.1', '42', '56', '69', '79', '33', '9',
       '.4', '51.5', '15', '5', '24', '7', '2', '6', '1', '58', '>50',
       '35', '86', '32.0', '93.2', '0', '27', '89.6', '23', '98', '19',
       '91.8', '>95', '57', '71.5', '78.3', '63', '1.5', '53.74', '59.5',
       '44', '42.5', '26', '3.5', '48', '26.3', '47', '88.5'],
      dtype=object)
```

In [55]:

```
# > and < will be changed to whole numbers less than or greater than.
clsm_cut['%.blasts.in.bm'] = clsm_cut['%.blasts.in.bm'].replace(['>50'], 51)
clsm_cut['%.blasts.in.bm'] = clsm_cut['%.blasts.in.bm'].replace(['>95'], 96)

clsm_cut['%.blasts.in.bm'].unique()
```

Out[55]:

```
array(['94', '80', '91', '97', '87', nan, '40', '75', '83', '95', '85',
       '90', '70', '92', '72', '68', '88', '36', '81', '93', '34', '77.5',
       '46', '65', '50', '76', '71', '60', '73', '55', '0.5', '30', '62',
       '18', '82', '28', '41', '64', '84', '21', '51', '17', '49.4', '32',
       '29', '25', '59.3', '66', '20', '52', '54', '22', '10', '12',
       '46.0', '13', '67', '39', '25.9', '45', '37', '78', '8', '3',
       '54.8', '74', '96', '4', '86.1', '42', '56', '69', '79', '33', '9',
       '.4', '51.5', '15', '5', '24', '7', '2', '6', '1', '58', 51, '35',
       '86', '32.0', '93.2', '0', '27', '89.6', '23', '98', '19', '91.8',
       96, '57', '71.5', '78.3', '63', '1.5', '53.74', '59.5', '44',
       '42.5', '26', '3.5', '48', '26.3', '47', '88.5'], dtype=object)
```

%.blasts.in.pb Attribute:

In [56]:

```
clsm_cut['%.blasts.in.pb'].unique()
```

Out[56]:

```
array(['97', '19', '99', '80', nan, '51', '30', '41', '84', '77', '75',
       '63', '60', '96', '66', '45', '93', '9', '82', '15', '33', '0',
       '13', '94', '89', '83', '>90', '78', '72', '59', '32', '6', '29',
       '24', '64', '57', '52', '2.1', '<5', '17', '22', '5', '47', '56',
       '25', '23', '42', '65', '71', '8', '3.5', '66.3', '95', '44', '10',
       '28.6', '18', '58', '67', '40', '92', '54', '1.0', '2', '20', '28',
       '35', '85', '1', '42.4', '16', '49.1', '14', '88', '46', '7',
       '0.5', '79', '26', '87', '20.4', '68', '48', '5.3', '61', '90',
       '17.4', '57.4', '43.8', '50', '37', '4', '3', '12', '81', '11',
       '90.5', '"""rare"""', '90.2', '55', '12.0', 'rare', '39', '31.0',
       '86', '47.4', '27.4', '39.6', '83.0', '12.9', '5.0', '15.4', '9.5',
       '62', '64.6', '27.8', '69.14', '52.2', '91', '67.25', '49', '23.7',
       '48.6', '98', '74.8', '2.6', '43', '29.6', '47.5', '38', '2.5',
       '25.2', '3.56', '70', '99.2', '73', '26.7', '38.5', '7.7', '74',
       '93.3', '12.1', '11.2', '92.9', '98.4', '6.8', '10.5', '53', '3.1',
       '28.9', '72.9', '40.2', '31', '3.3', '42.1', '11.5', '77.8', '3.8',
       '59.5', '21.7', '53.2'], dtype=object)
```

In [57]:

```
#%.Blasts.in.PB attribute has 1 "rare" record with no flt3 nor npm1 input. This will be changed to NAN
clsm_cut['%.blasts.in.pb'] = clsm_cut['%.blasts.in.pb'].replace(['"""rare"""'], np.nan)
clsm_cut['%.blasts.in.pb'] = clsm_cut['%.blasts.in.pb'].replace(['rare'], np.nan)
# > and < will be changed to whole numbers less than or greater than.
clsm_cut['%.blasts.in.pb'] = clsm_cut['%.blasts.in.pb'].replace(['<5'], 4)
clsm_cut['%.blasts.in.pb'] = clsm_cut['%.blasts.in.pb'].replace(['>90'], 91)

clsm_cut['%.blasts.in.pb'].unique()
```

Out[57]:

```
array(['97', '19', '99', '80', nan, '51', '30', '41', '84', '77', '75',
       '63', '60', '96', '66', '45', '93', '9', '82', '15', '33', '0',
       '13', '94', '89', '83', 91, '78', '72', '59', '32', '6', '29',
       '24', '64', '57', '52', '2.1', 4, '17', '22', '5', '47', '56',
       '25', '23', '42', '65', '71', '8', '3.5', '66.3', '95', '44', '10',
       '28.6', '18', '58', '67', '40', '92', '54', '1.0', '2', '20', '28',
       '35', '85', '1', '42.4', '16', '49.1', '14', '88', '46', '7',
       '0.5', '79', '26', '87', '20.4', '68', '48', '5.3', '61', '90',
       '17.4', '57.4', '43.8', '50', '37', '4', '3', '12', '81', '11',
       '90.5', '90.2', '55', '12.0', '39', '31.0', '86', '47.4', '27.4',
       '39.6', '83.0', '12.9', '5.0', '15.4', '9.5', '62', '64.6', '27.8',
       '69.14', '52.2', '91', '67.25', '49', '23.7', '48.6', '98', '74.8',
       '2.6', '43', '29.6', '47.5', '38', '2.5', '25.2', '3.56', '70',
       '99.2', '73', '26.7', '38.5', '7.7', '74', '93.3', '12.1', '11.2',
       '92.9', '98.4', '6.8', '10.5', '53', '3.1', '28.9', '72.9', '40.2',
       '31', '3.3', '42.1', '11.5', '77.8', '3.8', '59.5', '21.7', '53.2'],
      dtype=object)
```

From Categorical to Numerical

Transform %.blasts.in.bm and %.blasts.in.pb from object to float:

In [58]:

```
clsm_cut['%.blasts.in.bm'] = clsm_cut['%.blasts.in.bm'].astype(float)
clsm_cut['%.blasts.in.pb'] = clsm_cut['%.blasts.in.pb'].astype(float)


clsm_cut.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 18 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   labid                       672 non-null    object 
 1   patientid                   672 non-null    int64  
 2   consensus_sex               672 non-null    object 
 3   inferred_ethnicity          670 non-null    object 
 4   isrelapse                   672 non-null    bool   
 5   istransformed               672 non-null    bool   
 6   priormalignancynonmyeloid  672 non-null    object 
 7   priormds                    672 non-null    object 
 8   priormdsmpn                 672 non-null    object 
 9   priormpn                    672 non-null    object 
 10  eln2017                     672 non-null    object 
 11  dxatspecimenacquisition     672 non-null    object 
 12  vitalstatus                 672 non-null    object 
 13  overallsurvival             607 non-null    float64
 14  %.blasts.in.bm              459 non-null    float64
 15  %.blasts.in.pb              448 non-null    float64
 16  flt3-itd                    670 non-null    object 
 17  npm1                        669 non-null    object 
dtypes: bool(2), float64(3), int64(1), object(12)
memory usage: 85.4+ KB
```

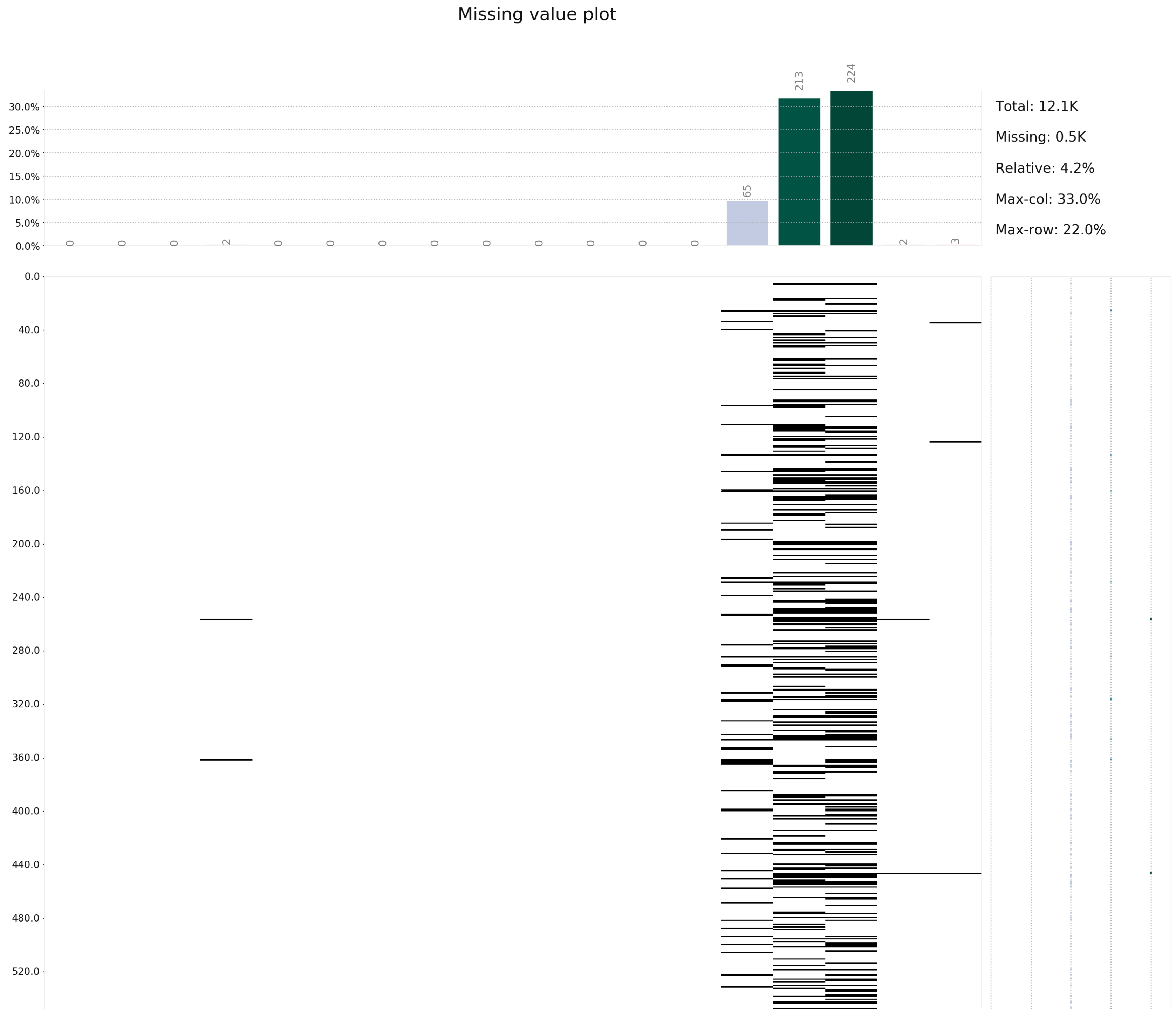
clsm_cut Identify Missing Values

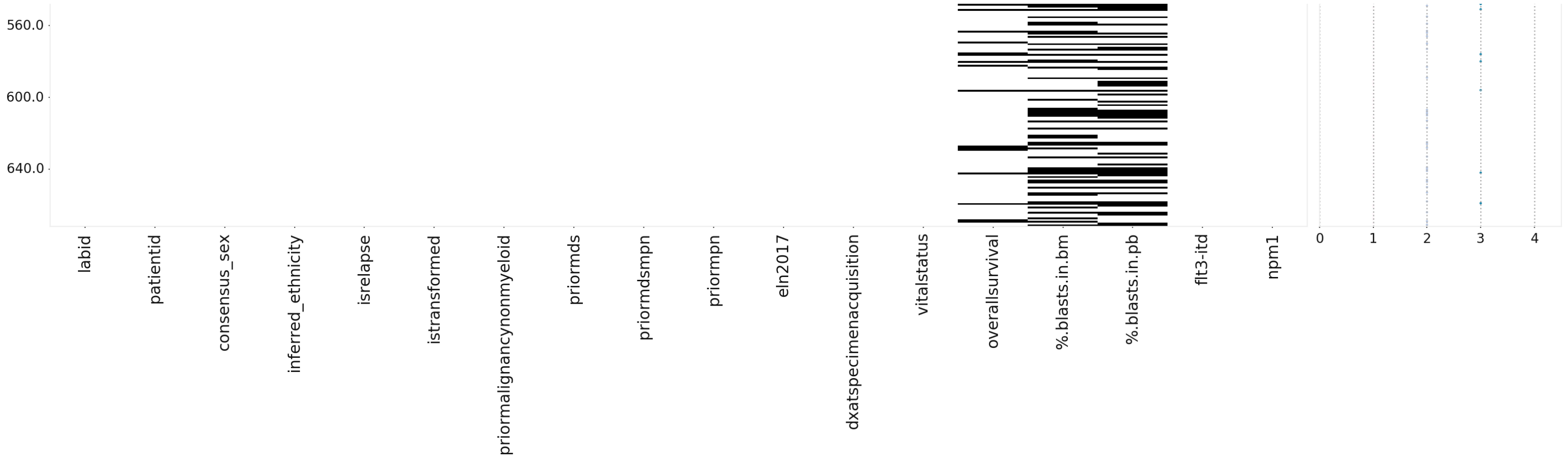
In [59]:

```
klib.missingval_plot(clsm_cut)
```

Out[59]:

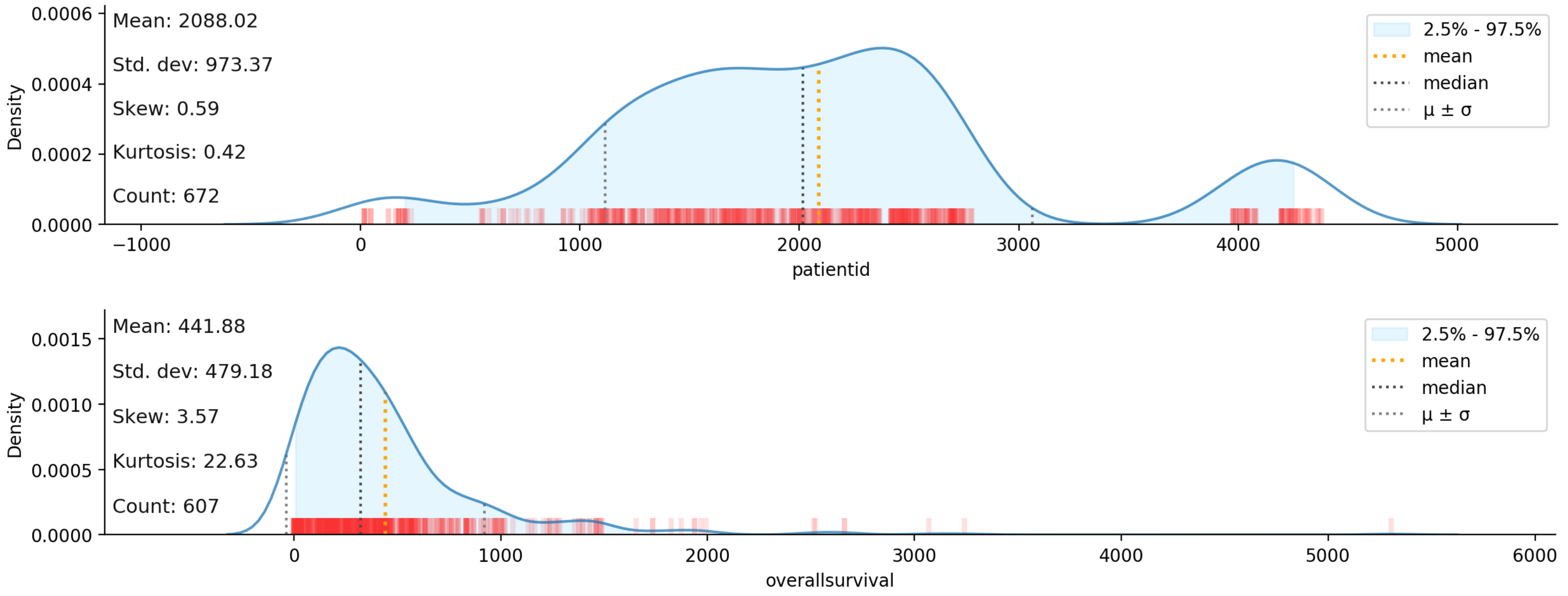
```
GridSpec(6, 6)
```

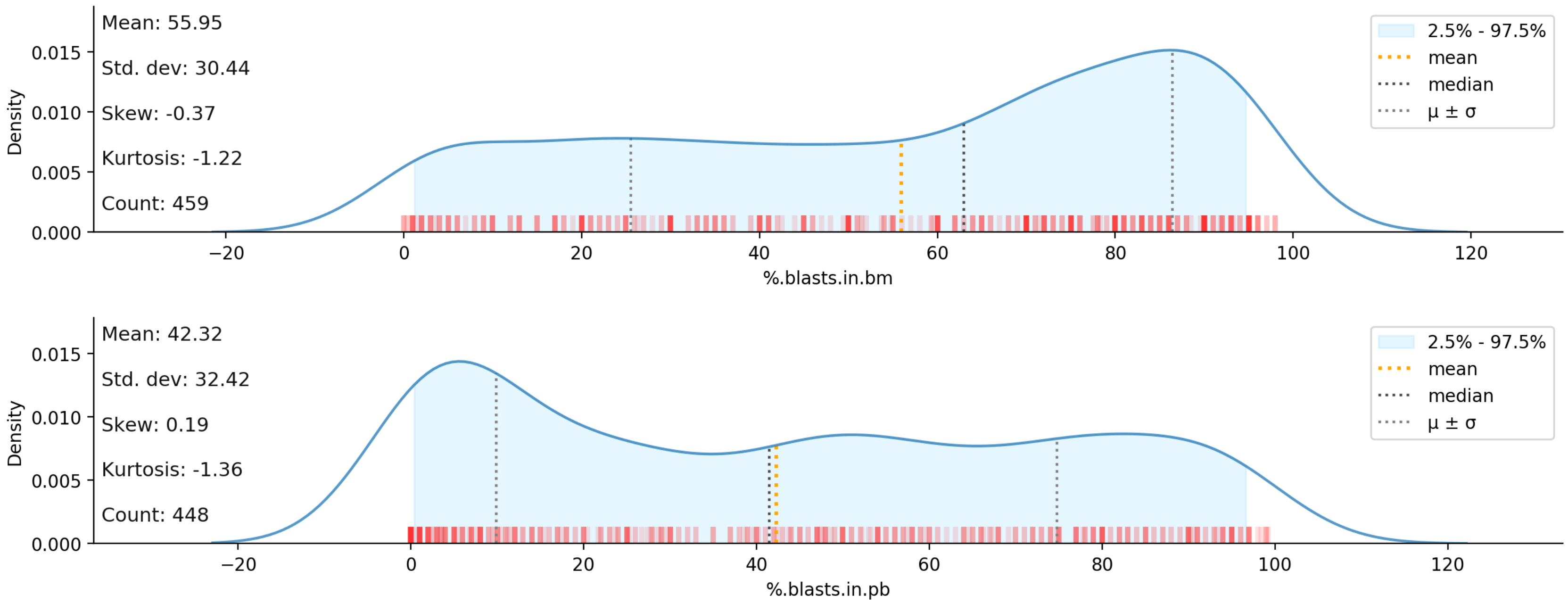

Missing value plot
0
0
0
2
0
0
0
0
0
0
0
0
0
65
213
224
2
3
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Total: 12.1K
Missing: 0.5K
Relative: 4.2%
Max-col: 33.0%
Max-row: 22.0%
0.0
40.0
80.0
120.0
160.0
200.0
240.0
280.0
320.0
360.0
400.0
440.0
480.0
520.0



Replace Missing Values

```
In [60]: klib.dist_plot(clsm_cut)

Out[60]: <matplotlib.axes._subplots.AxesSubplot at 0x7fce464e0750>
```





```
In [61]: clsm_cut.describe()
```

Out[61]:

	patientid	overallsurvival	%.blasts.in.bm	%.blasts.in.pb
count	672.000000	607.000000	459.000000	448.000000
mean	2088.020833	441.881384	55.949325	42.316629
std	973.372734	479.180429	30.440925	32.418249
min	17.000000	-1.000000	0.000000	0.000000
25%	1450.750000	167.000000	30.000000	10.000000
50%	2016.000000	323.000000	63.000000	41.500000
75%	2501.500000	555.000000	83.000000	72.000000
max	4380.000000	5305.000000	98.000000	99.200000

In [62]:

```
#From distibution, skewness suggest median is the best representation.
clsm_cut['overallsurvival'] = clsm_cut['overallsurvival'].fillna(clsm_cut['overallsurvival'].median())
clsm_cut['%.blasts.in.bm'] = clsm_cut['%.blasts.in.bm'].fillna(clsm_cut['%.blasts.in.bm'].median())
clsm_cut['%.blasts.in.pb'] = clsm_cut['%.blasts.in.pb'].fillna(clsm_cut['%.blasts.in.pb'].median())
```

In [63]:

```
#Replace categorical NaN with unknown
clsm_cut = clsm_cut.replace(np.nan, 'unknown', regex=True)
```

In [64]:

```
#Determine mode of inferred_ethnicity:
clsm_cut['inferred_ethnicity'].mode()
```

Out[64]:

```
0    White
dtype: object
```

In [65]:

```
#In inferred_ethnicity, replace mode of unknown to white:
clsm_cut['inferred_ethnicity'] = clsm_cut['inferred_ethnicity'].replace(['unknown'], 'white')

clsm_cut['inferred_ethnicity'].unique()
```

Out[65]:

```
array(['White', 'HispNative', 'AdmixedBlack', 'Asian', 'Black',
       'AdmixedAsian', 'white', 'AdmixedWhite', 'AdmixedHispNative'],
      dtype=object)
```

In [66]:

```
#Determine mode of flt3-itd:
clsm_cut['flt3-itd'].mode()
```

Out[66]:

```
0    negative
dtype: object
```

In [67]:

```
#In flt3-itd, replace mode of unknown to negative:
clsm_cut['flt3-itd'] = clsm_cut['flt3-itd'].replace(['unknown'], 'negative')

clsm_cut['flt3-itd'].unique()
```

Out[67]:

```
array(['positive', 'negative'], dtype=object)
```

In [68]:

```
#Determine mode of npm1:
clsm_cut['npm1'].mode()
```

Out[68]:

```
0    negative
dtype: object
```

In [69]:

```
#In npm1, replace mode of unknown to negative:
clsm_cut['npm1'] = clsm_cut['npm1'].replace(['unknown'], 'negative')

clsm_cut['npm1'].unique()
```

Out[69]:

```
array(['positive', 'negative'], dtype=object)
```

In [70]:

```
klib.missingval_plot(clsm_cut)
```

```
No missing values found in the dataset.
```

In [71]:

```
clsm_cut.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 18 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   labid                       672 non-null    object 
 1   patientid                   672 non-null    int64  
 2   consensus_sex               672 non-null    object 
 3   inferred_ethnicity          672 non-null    object 
 4   isrelapse                   672 non-null    bool   
 5   istransformed               672 non-null    bool   
 6   priormalignancynonmyeloid   672 non-null    object 
 7   priormds                    672 non-null    object 
 8   priormdsmpn                 672 non-null    object 
 9   priormpn                    672 non-null    object 
 10  eln2017                     672 non-null    object 
 11  dxatspecimenacquisition     672 non-null    object 
 12  vitalstatus                 672 non-null    object 
 13  overallsurvival             672 non-null    float64
 14  %.blasts.in.bm              672 non-null    float64
 15  %.blasts.in.pb              672 non-null    float64
 16  flt3-itd                    672 non-null    object 
 17  npm1                        672 non-null    object 
dtypes: bool(2), float64(3), int64(1), object(12)
memory usage: 85.4+ KB
```

Check for Duplicates

In [72]:

```
clsm_cut = clsm_cut.drop_duplicates(ignore_index=True)
clsm_cut.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 18 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   labid                       672 non-null    object 
 1   patientid                   672 non-null    int64  
 2   consensus_sex               672 non-null    object 
 3   inferred_ethnicity          672 non-null    object 
 4   isrelapse                   672 non-null    bool   
 5   istransformed               672 non-null    bool   
 6   priormalignancynonmyeloid  672 non-null    object 
 7   priormds                    672 non-null    object 
 8   priormdsmpn                 672 non-null    object 
 9   priormpn                    672 non-null    object 
 10  eln2017                     672 non-null    object 
 11  dxatspecimenacquisition     672 non-null    object 
 12  vitalstatus                 672 non-null    object 
 13  overallsurvival             672 non-null    float64
 14  %.blasts.in.bm              672 non-null    float64
 15  %.blasts.in.pb              672 non-null    float64
 16  flt3-itd                    672 non-null    object 
 17  npm1                        672 non-null    object 
dtypes: bool(2), float64(3), int64(1), object(12)
memory usage: 85.4+ KB
```

Create Target Variable

In [73]:

```
clsm_cut['dxatspecimenacquisition'].value_counts()
```

Out[73]:

```
ACUTE MYELOID LEUKAEMIA (AML) AND RELATED PRECURSOR NEOPLASMS    646
MYELODYSPLASTIC SYNDROMES                                         15
MYELODYSPLASTIC/MYELOPROLIFERATIVE NEOPLASMS                       4
ACUTE LEUKAEMIAS OF AMBIGUOUS LINEAGE                              3
MYELOPROLIFERATIVE NEOPLASMS                                       3
MATURE B-CELL NEOPLASMS                                            1
Name: dxatspecimenacquisition, dtype: int64
```

In [74]:

```
#create column for AML detected
clsm_cut['AML_detected'] = ['yes' if x == 'ACUTE MYELOID LEUKAEMIA (AML) AND RELATED PRECURSOR NEOPLASMS'
                            else 'no' for x in clsm_cut['dxatspecimenacquisition']]
```

In [75]:

```
clsm_cut.head()
```

Out[75]:

	labid	patientid	consensus_sex	inferred_ethnicity	isrelapse	istransformed	priormalignancynonmyeloid	priormds	priormdsmpn	p
0	09-00705	163	Male	White	False	False	n	n	n	
1	10-00136	174	Male	White	False	False	n	n	n	
2	10-00172	175	Female	White	False	False	n	n	n	
3	10-00507	45	Female	White	False	False	n	n	n	
4	10-00542	174	Male	White	True	False	n	n	n	

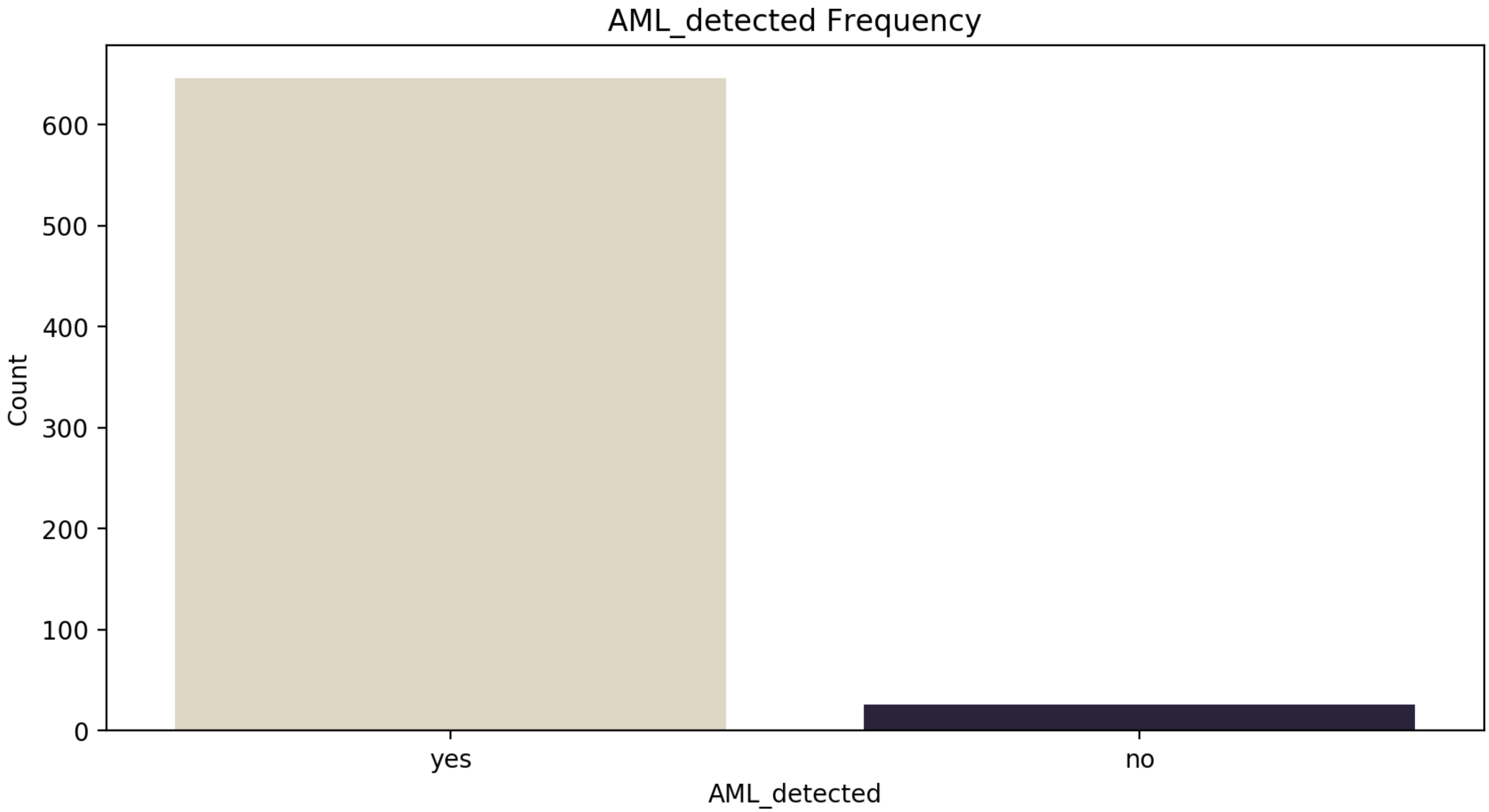
Data Exploration

In [77]:

```
sns.countplot(x=clsm_cut["AML_detected"], palette = "ch:s=-.2,r=.6")
plt.gcf().set_size_inches(10, 5)
plt.xlabel('AML_detected')
plt.ylabel('Count')
plt.title("AML_detected Frequency")
```

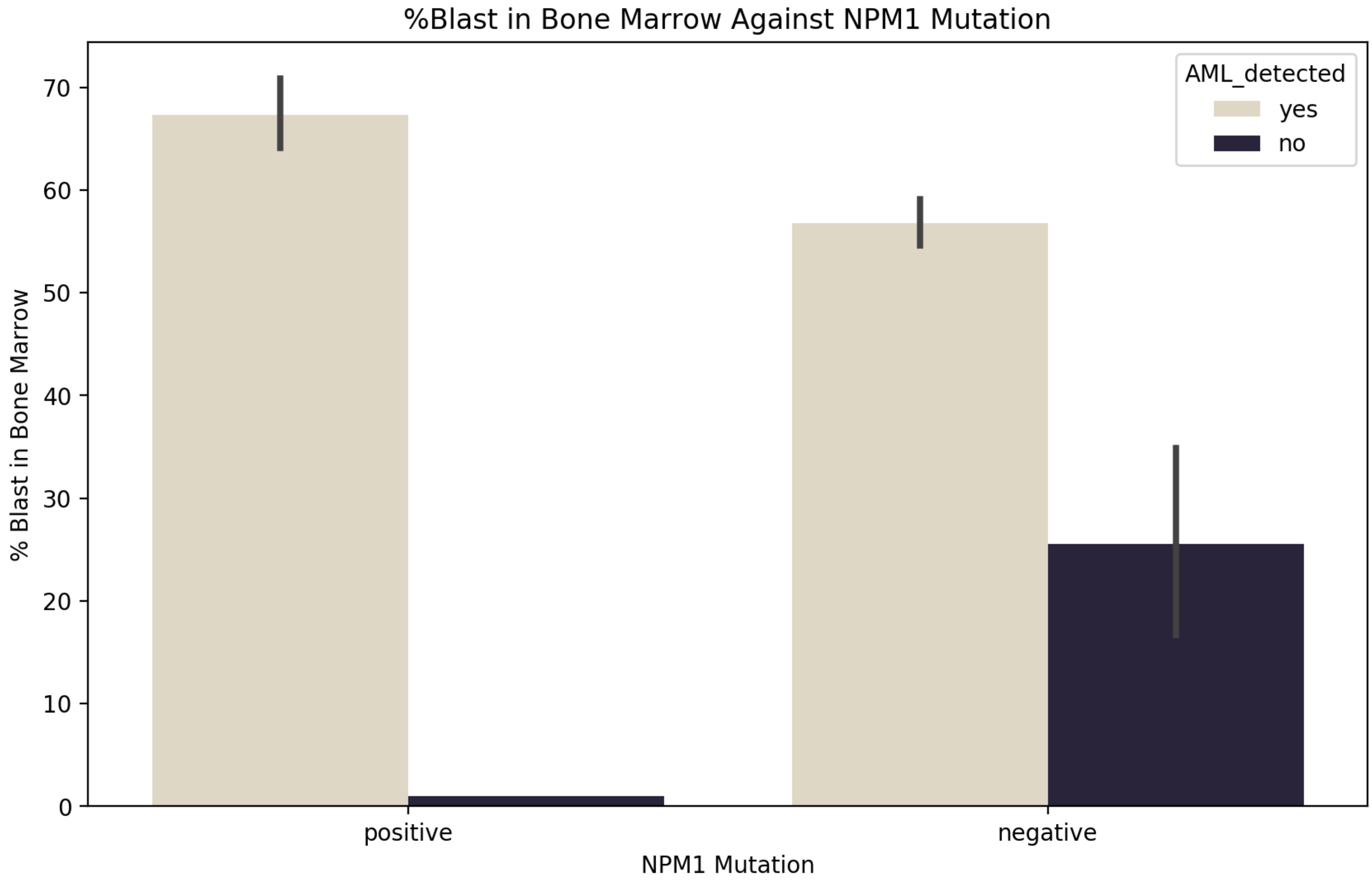
Out[77]:

```
Text(0.5, 1.0, 'AML_detected Frequency')
```



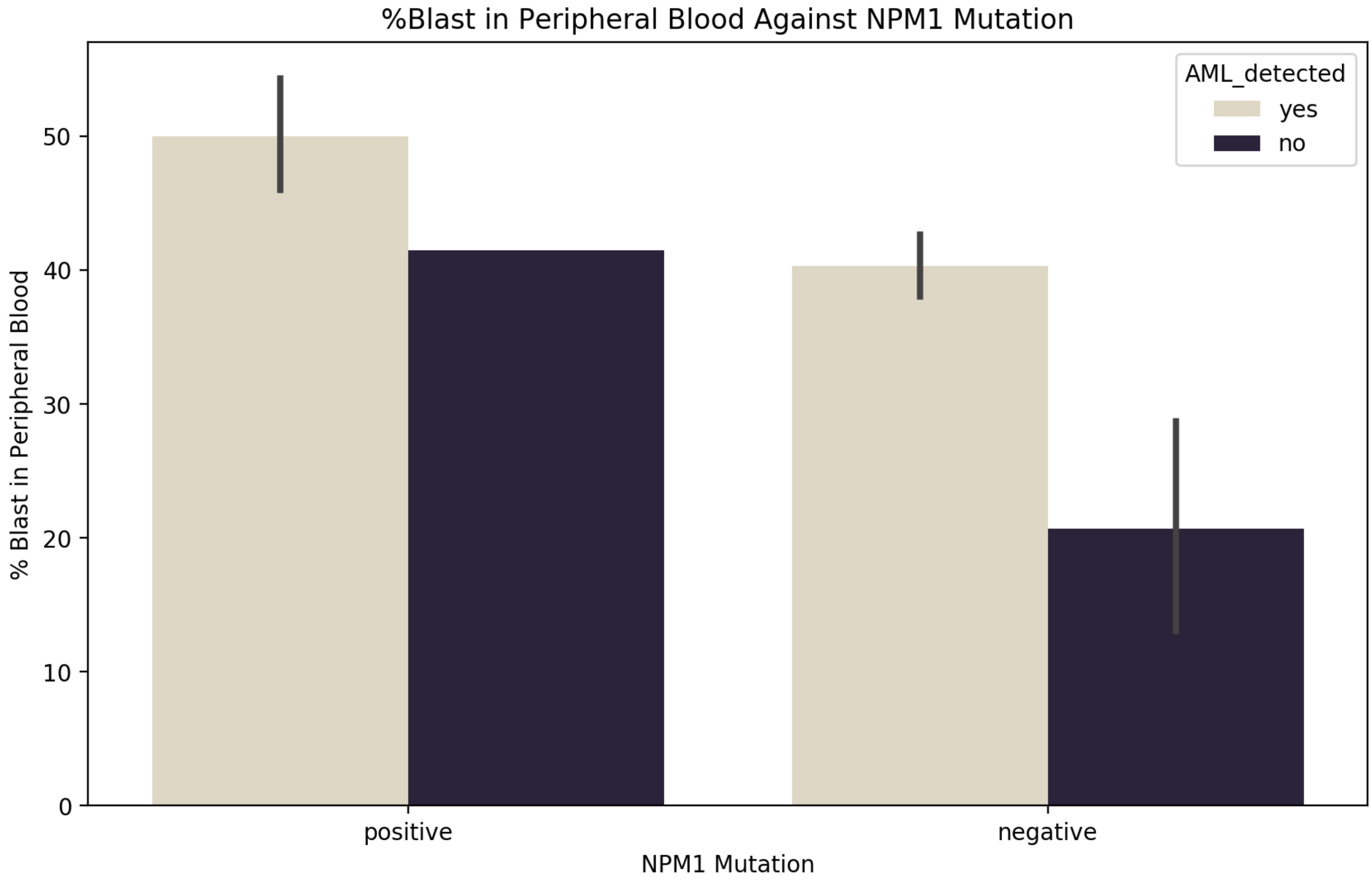
```
In [78]: sns.barplot(data= clsm_cut,x = 'npm1', y = '%.blasts.in.bm',
                     hue = 'AML_detected', palette = "ch:s=-.2,r=.6")
         plt.gcf().set_size_inches(10, 6)
         plt.xlabel('NPM1 Mutation')
         plt.ylabel('% Blast in Bone Marrow')
         plt.legend(loc='upper right', title = 'AML_detected')
         plt.title("%Blast in Bone Marrow Against NPM1 Mutation")
```

```
Out[78]: Text(0.5, 1.0, '%Blast in Bone Marrow Against NPM1 Mutation')
```



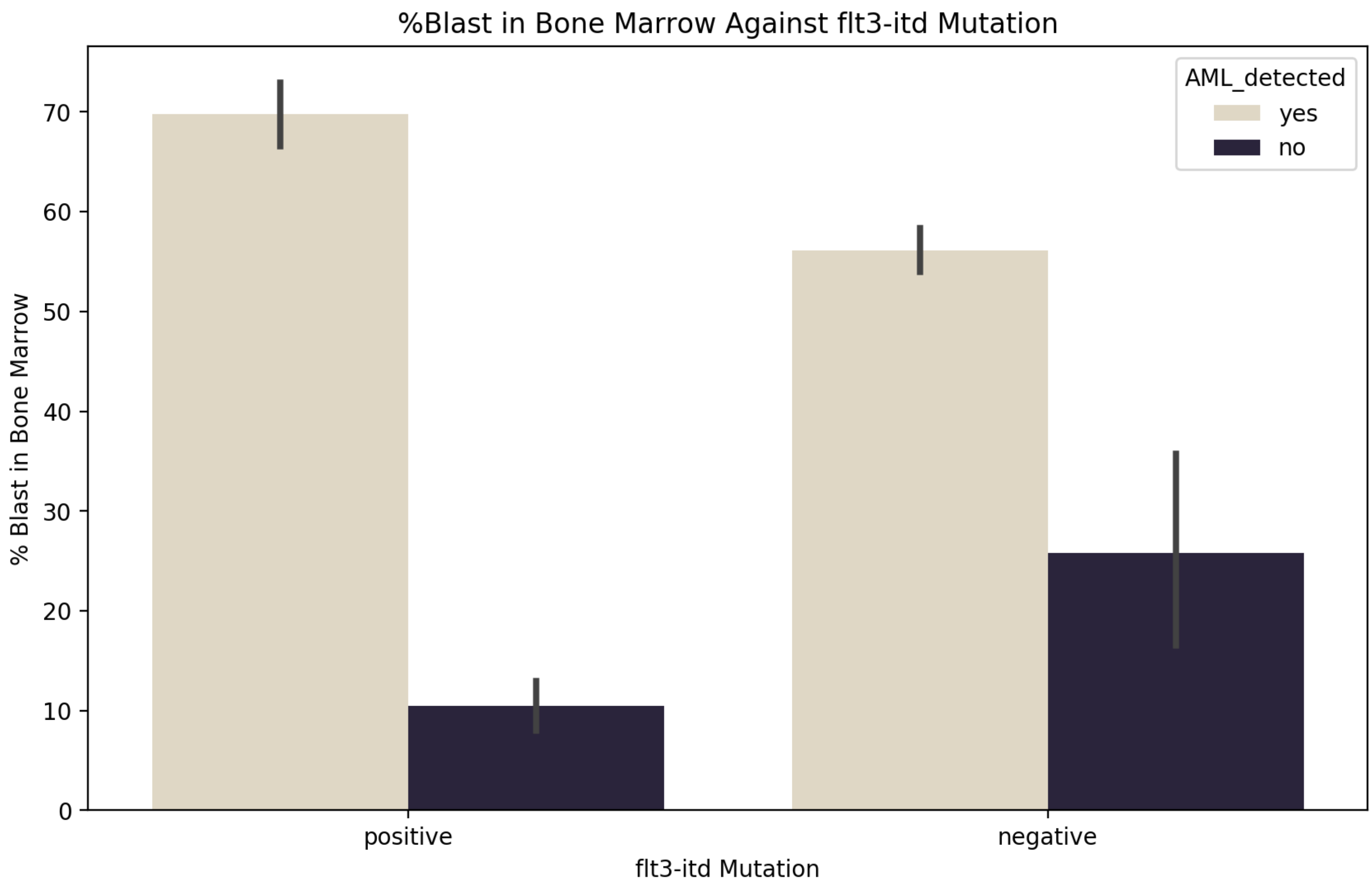
```
In [79]: sns.barplot(data= clsm_cut,x = 'npm1', y = '%.blasts.in.pb',
                     hue = 'AML_detected', palette = "ch:s=-.2,r=.6")
         plt.gcf().set_size_inches(10, 6)
         plt.xlabel('NPM1 Mutation')
         plt.ylabel('% Blast in Peripheral Blood')
         plt.legend(loc='upper right', title = 'AML_detected')
         plt.title("%Blast in Peripheral Blood Against NPM1 Mutation")
```

```
Out[79]: Text(0.5, 1.0, '%Blast in Peripheral Blood Against NPM1 Mutation')
```



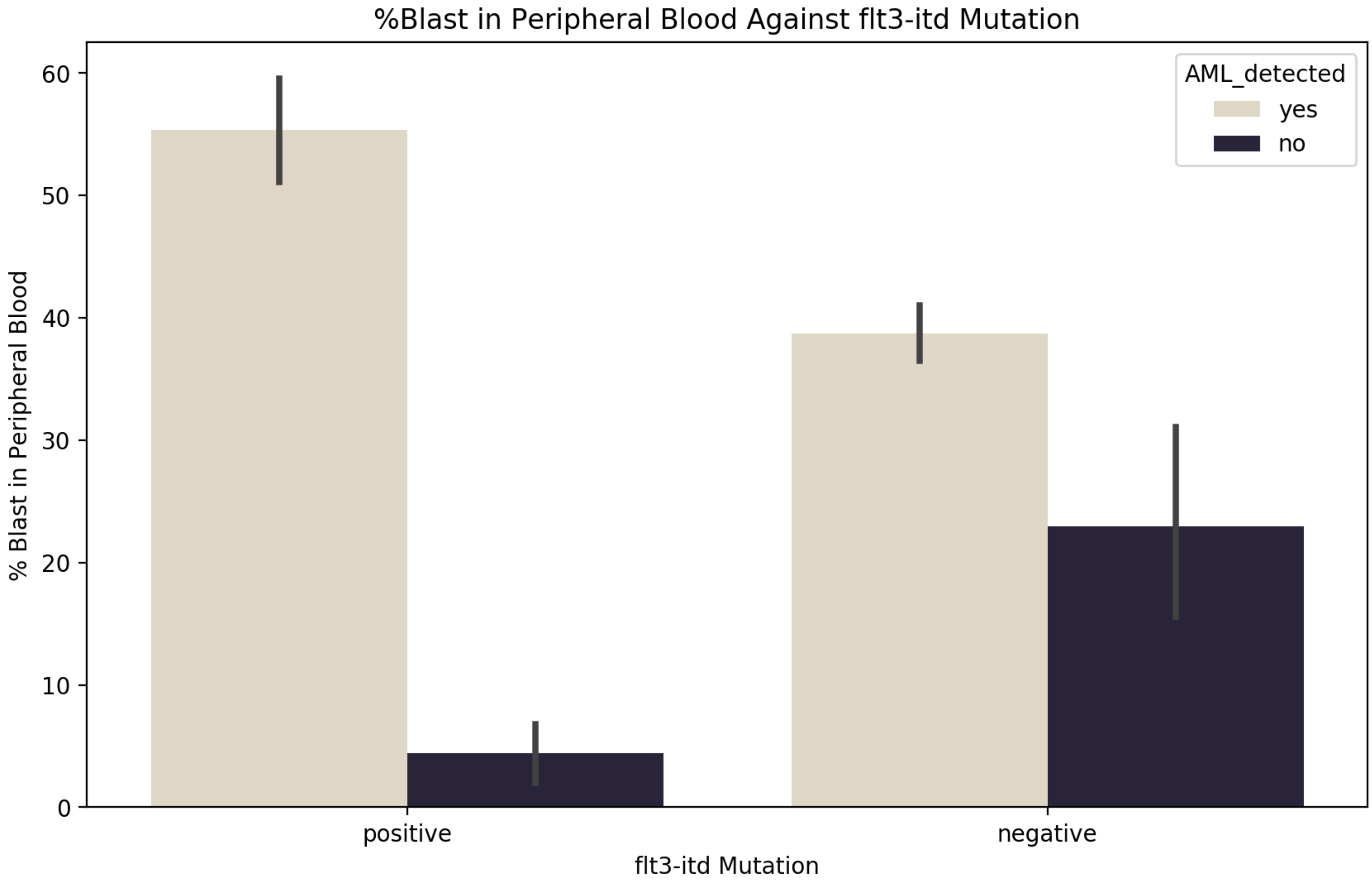
In [80]:
```
sns.barplot(data= clsm_cut,x = 'flt3-itd', y = '%.blasts.in.bm',
            hue = 'AML_detected', palette = "ch:s=-.2,r=.6")
plt.gcf().set_size_inches(10, 6)
plt.xlabel('flt3-itd Mutation')
plt.ylabel('% Blast in Bone Marrow')
plt.legend(loc='upper right', title = 'AML_detected')
plt.title("%Blast in Bone Marrow Against flt3-itd Mutation")
```

Out[80]:
```
Text(0.5, 1.0, '%Blast in Bone Marrow Against flt3-itd Mutation')
```

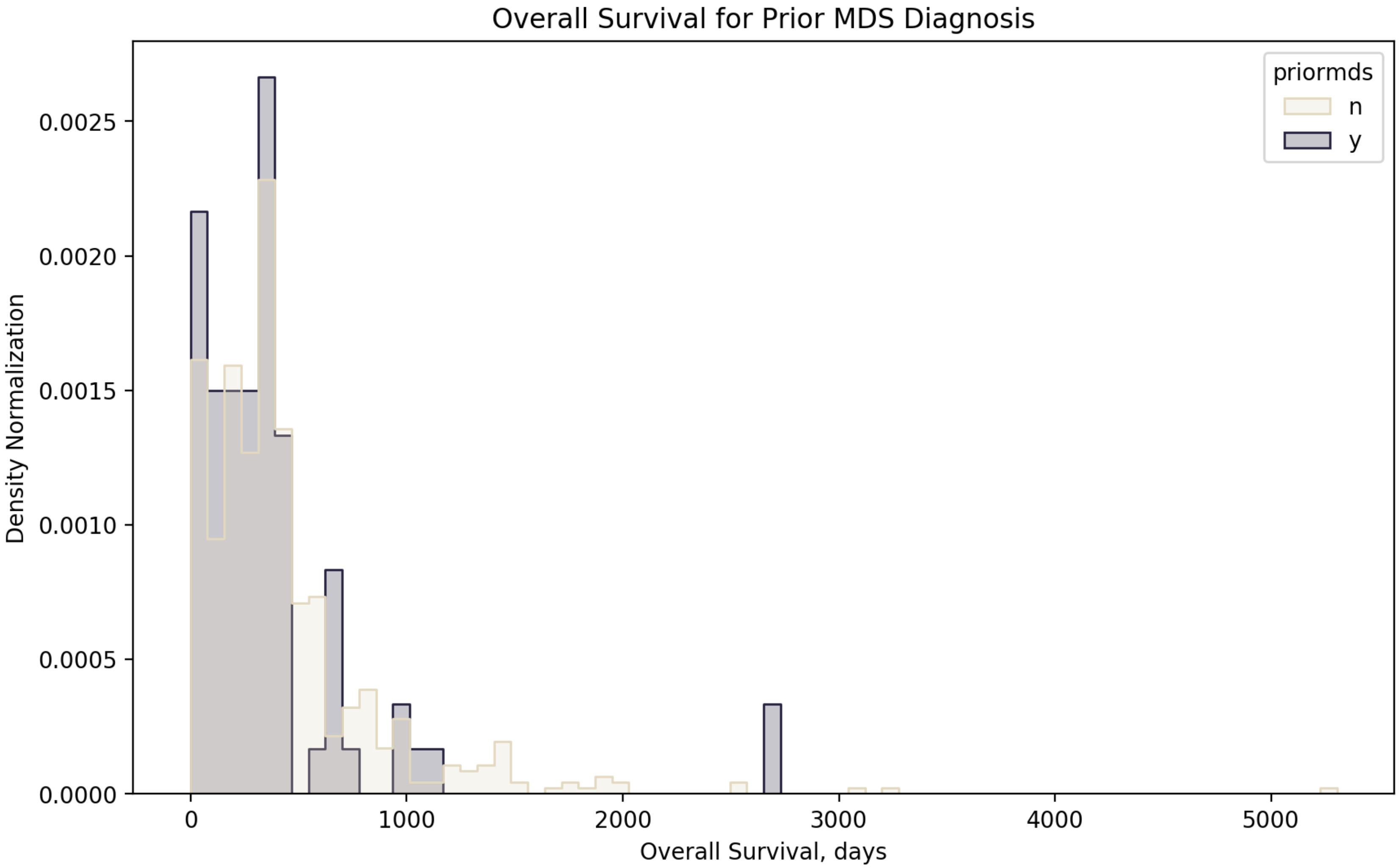


```
In [81]: sns.barplot(data= clsm_cut,x = 'flt3-itd', y = '%.blasts.in.pb',
                     hue = 'AML_detected', palette = "ch:s=-.2,r=.6")
         plt.gcf().set_size_inches(10, 6)
         plt.xlabel('flt3-itd Mutation')
         plt.ylabel('% Blast in Peripheral Blood')
         plt.legend(loc='upper right', title = 'AML_detected')
         plt.title("%Blast in Peripheral Blood Against flt3-itd Mutation")
```

```
Out[81]: Text(0.5, 1.0, '%Blast in Peripheral Blood Against flt3-itd Mutation')
```

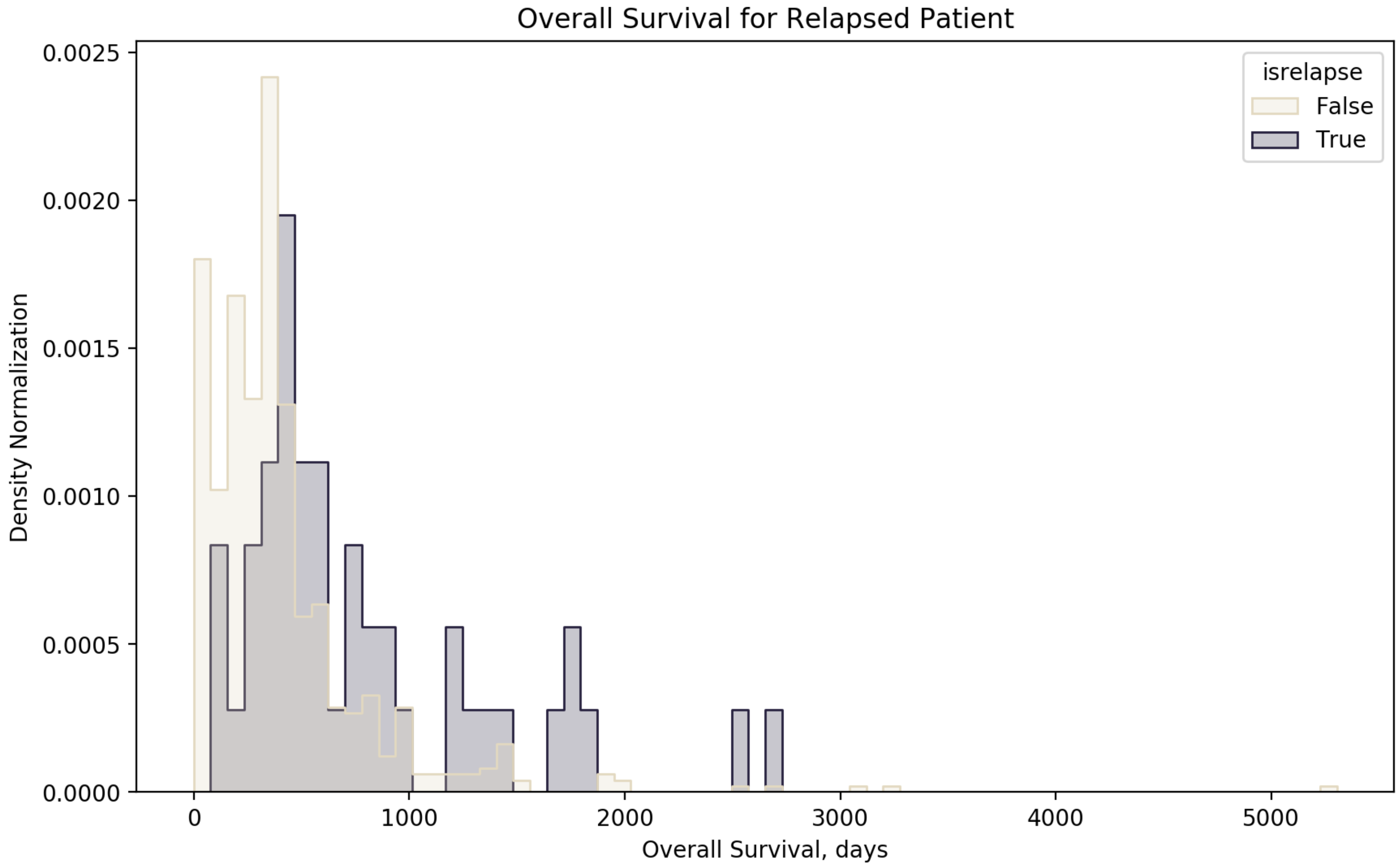


```
In [82]: sns.histplot(data=clsm_cut, x="overallsurvival", hue="priormds", element="step", palette = "ch:s=-.2,r=.6",
                      stat="density", common_norm=False)
         plt.gcf().set_size_inches(10, 6)
         plt.xlabel('Overall Survival, days')
         plt.ylabel('Density Normalization')
         #plt.legend(loc='upper right', title = 'Prior MDS Diagnosis')
         plt.title("Overall Survival for Prior MDS Diagnosis")

Out[82]: Text(0.5, 1.0, 'Overall Survival for Prior MDS Diagnosis')
```



```
In [83]: sns.histplot(data=clsm_cut, x="overallsurvival", hue="isrelapse", element="step", palette = "ch:s=-.2,r=.6",
                      stat="density", common_norm=False)
         plt.gcf().set_size_inches(10, 6)
         plt.xlabel('Overall Survival, days')
         plt.ylabel('Density Normalization')
         #plt.legend(loc='upper right', title = 'Prior MDS Diagnosis')
         plt.title("Overall Survival for Relapsed Patient")
```

```
Out[83]: Text(0.5, 1.0, 'Overall Survival for Relapsed Patient')
```



New Dataframe For SageMaker JumpStart Regression Model

Transform select categorical attributes to numerical:

In [84]:

```
#AML_detected
clsm_cut['AML_detected'].replace(['no', 'yes'],
                        [0, 1], inplace=True)

#npm1
clsm_cut['npm1'].replace(['negative', 'positive'],
                        [0, 1], inplace=True)

#flt3-itd
clsm_cut['flt3-itd'].replace(['negative', 'positive'],
                        [0, 1], inplace=True)

#priormalignancynonmyeloid
clsm_cut['priormalignancynonmyeloid'].replace(['n', 'y'],
                        [0, 1], inplace=True)

#priormds
clsm_cut['priormds'].replace(['y', 'n'],
                        [1, 0], inplace=True)

#priormdsmpn
clsm_cut['priormdsmpn'].replace(['n', 'y'],
                        [0, 1], inplace=True)

#priormpn
clsm_cut['priormpn'].replace(['n', 'y'],
                        [0, 1], inplace=True)

#isrelapse
clsm_cut['isrelapse'].replace(['False', 'True'],
                        [0, 1], inplace=True)

#istransformed
clsm_cut['istransformed'].replace(['True', 'False'],
                        [1, 0], inplace=True)
```

In [85]:

```
clsm_t = pd.DataFrame(clsm_cut[['AML_detected', 'npm1', 'flt3-itd', 'isrelapse', 'istransformed',
                                'priormalignancynonmyeloid', 'priormds', 'priormdsmpn', 'priormpn',
                                '%.blasts.in.pb', '%.blasts.in.bm', 'overallsurvival']])
```

```
In [86]: #Transform data type:
         clsm_t['npm1'] = clsm_cut['npm1'].astype(int)
         clsm_t['flt3-itd'] = clsm_cut['flt3-itd'].astype(int)

         clsm_t['isrelapse'] = clsm_cut['isrelapse'].astype(int)
         clsm_t['istransformed'] = clsm_cut['istransformed'].astype(int)
```

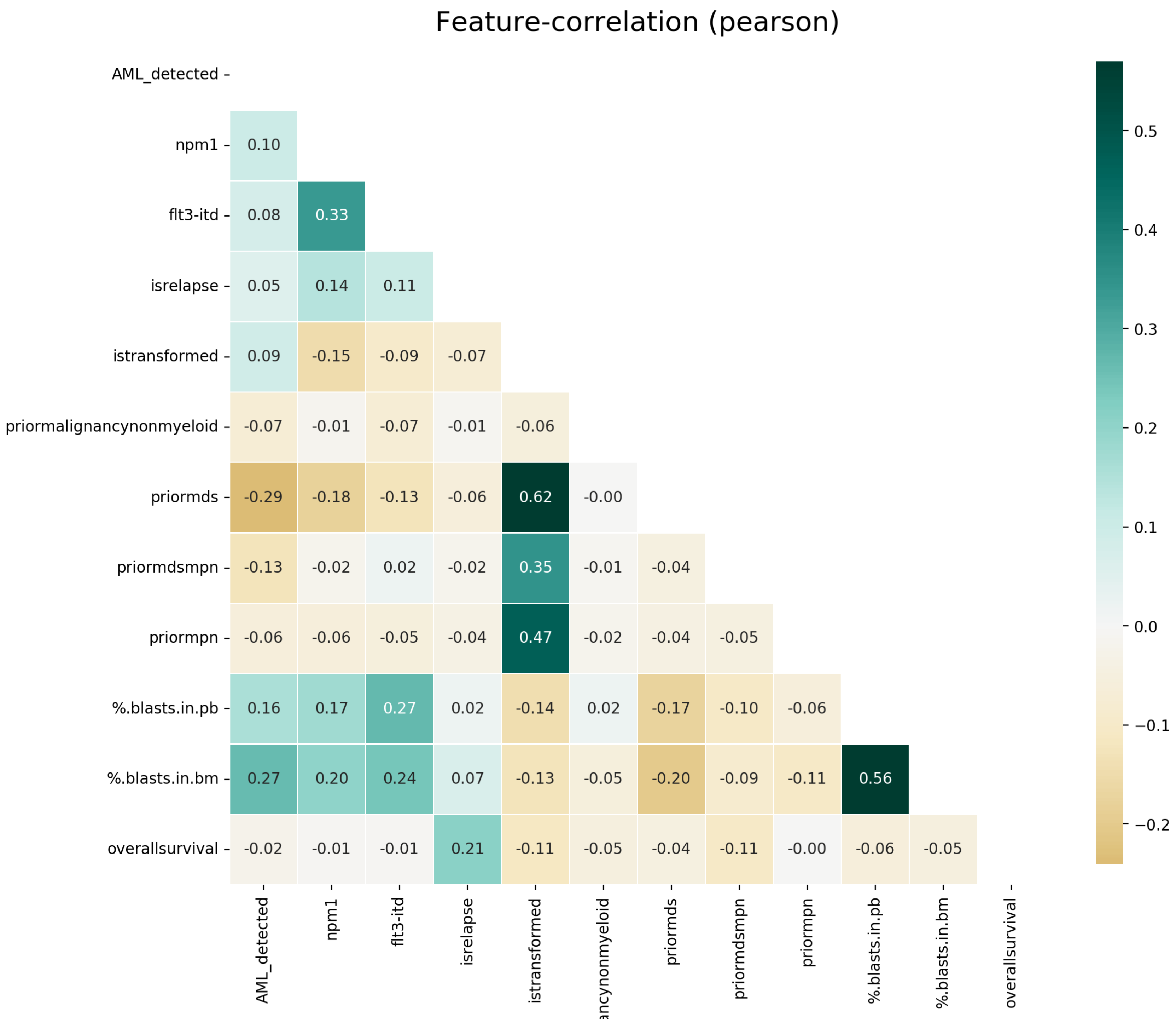
New clsm Dataframe Correlation Matrix

```
In [87]: clsm_t.corr()
```

Out[87]:

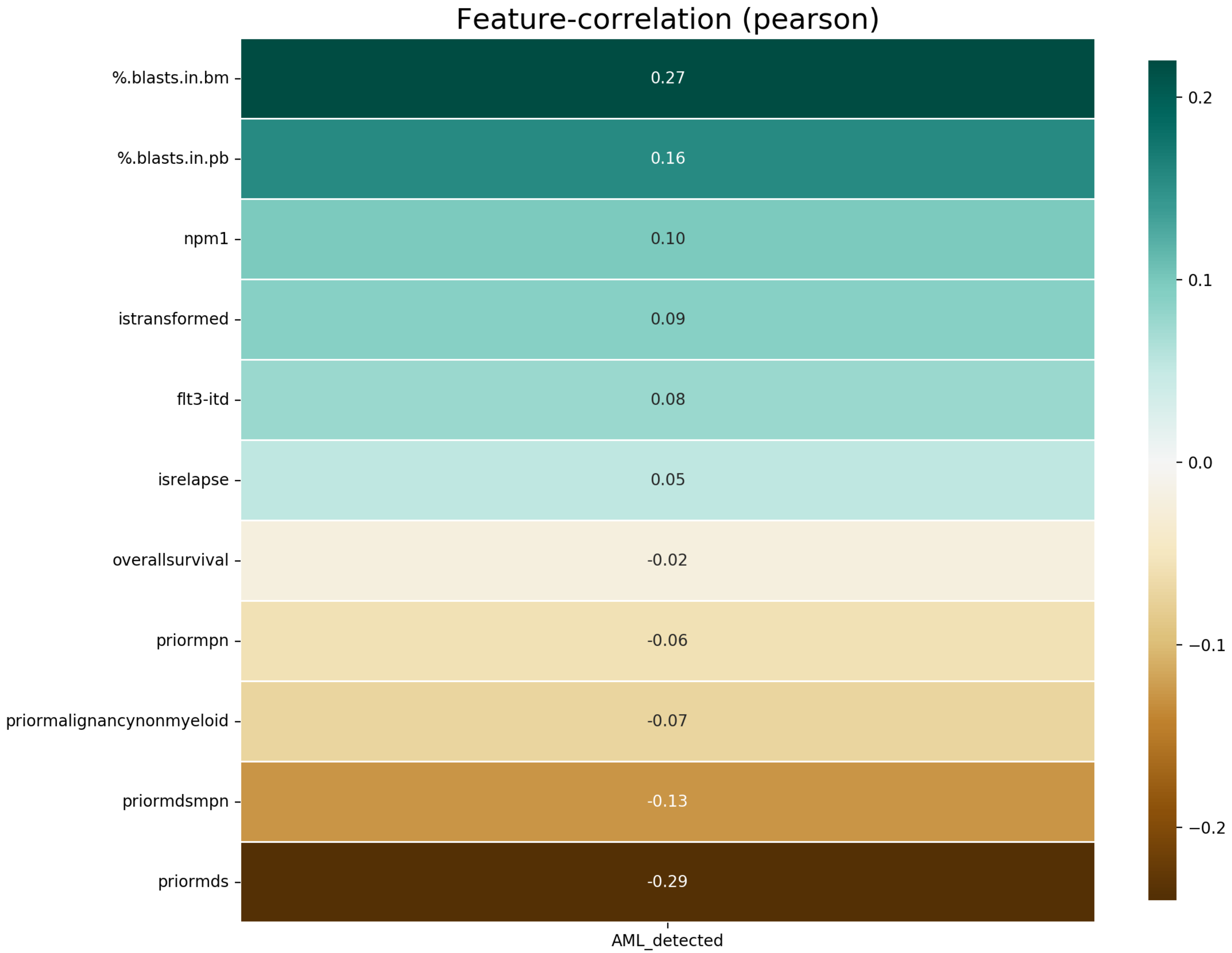
	AML_detected	npm1	flt3-itd	isrelapse	istransformed	priormalignancynonmyeloid	priormds	prio
AML_detected	1.000000	0.098997	0.077525	0.054383	0.089238	-7.245182e-02	-2.912038e-01	
npm1	0.098997	1.000000	0.333543	0.140481	-0.148233	-1.257739e-02	-1.771024e-01	
flt3-itd	0.077525	0.333543	1.000000	0.107818	-0.092782	-7.228395e-02	-1.272762e-01	
isrelapse	0.054383	0.140481	0.107818	1.000000	-0.072971	-9.623173e-03	-6.051426e-02	
istransformed	0.089238	-0.148233	-0.092782	-0.072971	1.000000	-5.562376e-02	6.200179e-01	
priormalignancynonmyeloid	-0.072452	-0.012577	-0.072284	-0.009623	-0.055624	1.000000e+00	-1.056121e-17	
priormds	-0.291204	-0.177102	-0.127276	-0.060514	0.620018	-1.056121e-17	1.000000e+00	
priormdsmpn	-0.127707	-0.019763	0.022049	-0.020414	0.346275	-9.820928e-03	-4.405654e-02	
priormpn	-0.057154	-0.059377	-0.054524	-0.037020	0.472862	-1.913898e-02	-4.227151e-02	
%.blasts.in.pb	0.155752	0.174675	0.271851	0.020293	-0.141930	2.215108e-02	-1.739113e-01	
%.blasts.in.bm	0.265724	0.201114	0.242420	0.069284	-0.128224	-5.402601e-02	-2.015171e-01	
overallsurvival	-0.022216	-0.006728	-0.008150	0.210147	-0.113017	-4.893419e-02	-4.466673e-02	

```
In [88]: klib.corr_plot(clsm_t)

Out[88]: <matplotlib.axes._subplots.AxesSubplot at 0x7fce4bac4950>
```


Feature-correlation (pearson)
AML_detected
npm1 0.10
flt3-itd 0.08 0.33
isrelapse 0.05 0.14 0.11
istransformed 0.09 -0.15 -0.09 -0.07
priormalignancynonmyeloid -0.07 -0.01 -0.07 -0.01 -0.06
priormds -0.29 -0.18 -0.13 -0.06 0.62 -0.00
priormdsmpn -0.13 -0.02 0.02 -0.02 0.35 -0.01 -0.04
priormpn -0.06 -0.06 -0.05 -0.04 0.47 -0.02 -0.04 -0.05
%.blasts.in.pb 0.16 0.17 0.27 0.02 -0.14 0.02 -0.17 -0.10 -0.06
%.blasts.in.bm 0.27 0.20 0.24 0.07 -0.13 -0.05 -0.20 -0.09 -0.11 0.56
overallsurvival -0.02 -0.01 -0.01 0.21 -0.11 -0.05 -0.04 -0.11 -0.00 -0.06 -0.05
AML_detected
npm1
flt3-itd
isrelapse
istransformed
ancynonmyeloid
priormds
priormdsmpn
priormpn
%.blasts.in.pb
%.blasts.in.bm
overallsurvival
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2

```
In [89]: klib.corr_plot(clsm_t, target='AML_detected')
```

```
Out[89]: <matplotlib.axes._subplots.AxesSubplot at 0x7fce46246bd0>
```



One-Hot encoding

In [90]:

```
clsm_t = pd.get_dummies(clsm_t, columns= ['npm1', 'flt3-itd', 'priormalignancynonmyeloid',
                                          'priormds', 'priormdsmpn', 'priormpn', 'isrelapse', 'istransformed'
```

In [91]:

```
clsm_t.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   AML_detected                 672 non-null    int64  
 1   %.blasts.in.pb               672 non-null    float64
 2   %.blasts.in.bm               672 non-null    float64
 3   overallsurvival              672 non-null    float64
 4   npm1_0                       672 non-null    uint8  
 5   npm1_1                       672 non-null    uint8  
 6   flt3-itd_0                   672 non-null    uint8  
 7   flt3-itd_1                   672 non-null    uint8  
 8   priormalignancynonmyeloid_0  672 non-null    uint8  
 9   priormalignancynonmyeloid_1  672 non-null    uint8  
 10  priormds_0                   672 non-null    uint8  
 11  priormds_1                   672 non-null    uint8  
 12  priormdsmpn_0                672 non-null    uint8  
 13  priormdsmpn_1                672 non-null    uint8  
 14  priormpn_0                   672 non-null    uint8  
 15  priormpn_1                   672 non-null    uint8  
 16  isrelapse_0                  672 non-null    uint8  
 17  isrelapse_1                  672 non-null    uint8  
 18  istransformed_0              672 non-null    uint8  
 19  istransformed_1              672 non-null    uint8  
dtypes: float64(3), int64(1), uint8(16)
memory usage: 31.6 KB
```

In [92]:

```
clsm_t.head()
```

Out[92]:

	AML_detected	%.blasts.in.pb	%.blasts.in.bm	overallsurvival	npm1_0	npm1_1	flt3-itd_0	flt3-itd_1	priormalignancynonmyeloid_0	priormalign
0	1	97.0	94.0	425.0	0	1	0	1	1	
1	1	19.0	80.0	419.0	1	0	0	1	1	
2	1	99.0	91.0	541.0	1	0	0	1	1	
3	1	97.0	97.0	511.0	0	1	0	1	1	
4	1	80.0	87.0	419.0	1	0	0	1	1	

Transform Headers

In [93]:

```
clsm_t = clsm_t.rename(columns={ '%.blasts.in.pb': 'Feature_1', '%.blasts.in.bm': 'Feature_2',
                                 'overallsurvival': 'Feature_3',
                                  'npm1_0': 'Feature_4', 'npm1_1': 'Feature_5',
                                 'flt3-itd_0': 'Feature_6', 'flt3-itd_1': 'Feature_7',
                                  'priormalignancynonmyeloid_0': 'Feature_8', 'priormalignancynonmyeloid_1'
                                  'priormds_0': 'Feature_10', 'priormds_1': 'Feature_11',
                                 'priormdsmpn_0': 'Feature_12', 'priormdsmpn_1': 'Feature_13',
                                  'priormpn_0': 'Feature_14', 'priormpn_1': 'Feature_15',
                                  'isrelapse_0': 'Feature_16', 'isrelapse_1': 'Feature_17',
                                 'istransformed_0': 'Feature_18', 'istransformed_1': 'Feature_19' })
```

In [94]:

```
clsm_t.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 20 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   AML_detected  672 non-null    int64  
 1   Feature_1     672 non-null    float64
 2   Feature_2     672 non-null    float64
 3   Feature_3     672 non-null    float64
 4   Feature_4     672 non-null    uint8  
 5   Feature_5     672 non-null    uint8  
 6   Feature_6     672 non-null    uint8  
 7   Feature_7     672 non-null    uint8  
 8   Feature_8     672 non-null    uint8  
 9   Feature_9     672 non-null    uint8  
 10  Feature_10    672 non-null    uint8  
 11  Feature_11    672 non-null    uint8  
 12  Feature_12    672 non-null    uint8  
 13  Feature_13    672 non-null    uint8  
 14  Feature_14    672 non-null    uint8  
 15  Feature_15    672 non-null    uint8  
 16  Feature_16    672 non-null    uint8  
 17  Feature_17    672 non-null    uint8  
 18  Feature_18    672 non-null    uint8  
 19  Feature_19    672 non-null    uint8  
dtypes: float64(3), int64(1), uint8(16)
memory usage: 31.6 KB
```

Save New Pre-Processed clsm Dataframe to S3

In [95]:

```
clsm_t.to_csv('clsm_t.csv')
```

In [96]:

```
#Manually upload into S3
!aws s3 ls s3://team4rawdatasets/CSV/Input/
```

```
                           PRE OHSU_BeatAML_ClinicalSummary/
                           PRE OpenCell_ProteinInteraction/
2023-03-21 01:19:41          0 
2023-04-12 05:44:22      30850 clsm_t.csv
```

In [97]:

```
from IPython.core.display import display, HTML

display(
    HTML(
        '<b>Review <a target="blank" href="https://s3.console.aws.amazon.com/s3/buckets/team4rawdatasets?region=us-ea
            region, account_id, region
        )
    )
)
```

Review S3 Output Bucket

Split the Data into Train, Test, and Validation sets

In [98]:

```
clsm_t.info()
```

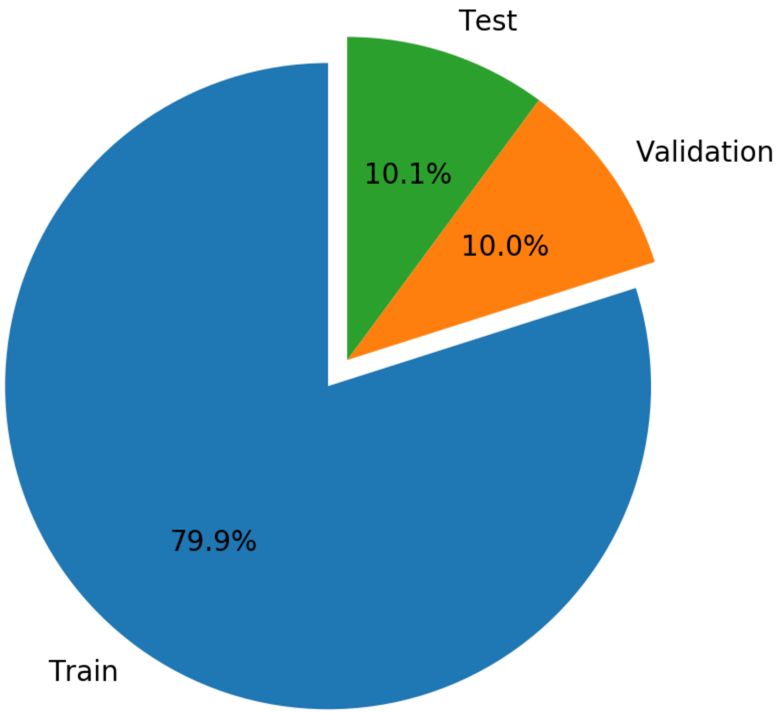
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 20 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   AML_detected  672 non-null    int64  
 1   Feature_1     672 non-null    float64
 2   Feature_2     672 non-null    float64
 3   Feature_3     672 non-null    float64
 4   Feature_4     672 non-null    uint8  
 5   Feature_5     672 non-null    uint8  
 6   Feature_6     672 non-null    uint8  
 7   Feature_7     672 non-null    uint8  
 8   Feature_8     672 non-null    uint8  
 9   Feature_9     672 non-null    uint8  
 10  Feature_10    672 non-null    uint8  
 11  Feature_11    672 non-null    uint8  
 12  Feature_12    672 non-null    uint8  
 13  Feature_13    672 non-null    uint8  
 14  Feature_14    672 non-null    uint8  
 15  Feature_15    672 non-null    uint8  
 16  Feature_16    672 non-null    uint8  
 17  Feature_17    672 non-null    uint8  
 18  Feature_18    672 non-null    uint8  
 19  Feature_19    672 non-null    uint8  
dtypes: float64(3), int64(1), uint8(16)
memory usage: 31.6 KB
```

In [99]:

```python
from sklearn.model_selection import train_test_split

# Split all data into 80% train and 20% holdout
clsm_train, clsm_holdout = train_test_split(clsm_t, test_size=0.20, random_state=42)

# Split holdout data into 50% validation and 50% test
clsm_validation, clsm_test = train_test_split(clsm_holdout, test_size=0.50, random_state=42)
```

In [100…

```
# Pie chart, where the slices will be ordered and plotted counter-clockwise:
labels = ["Train", "Validation", "Test"]
sizes = [len(clsm_train.index), len(clsm_validation.index), len(clsm_test.index)]
explode = (0.1, 0, 0)

fig1, ax1 = plt.subplots()

ax1.pie(sizes, explode=explode, labels=labels, autopct="%1.1f%%", startangle=90)

# Equal aspect ratio ensures that pie is drawn as a circle.
ax1.axis("equal")

plt.show()
```



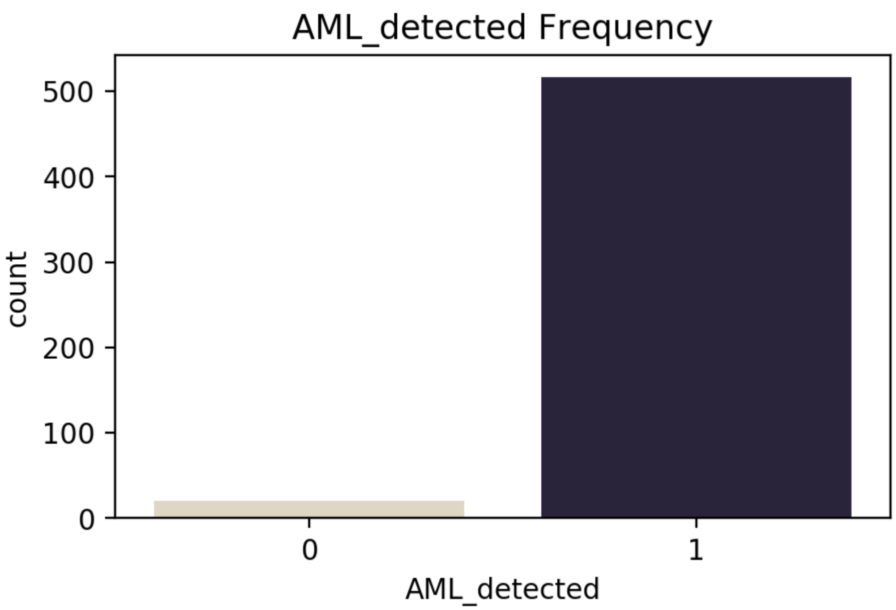
Show 80% Train Data Split

In [101…

```
clsm_train.shape
```

Out[101]:
```
(537, 20)
```

In [102…
```
sns.countplot(x=clsm_train["AML_detected"], palette = "ch:s=-.2,r=.6")
plt.xlabel('AML_detected')
plt.title('AML_detected Frequency')
plt.gcf().set_size_inches(5, 3)
```

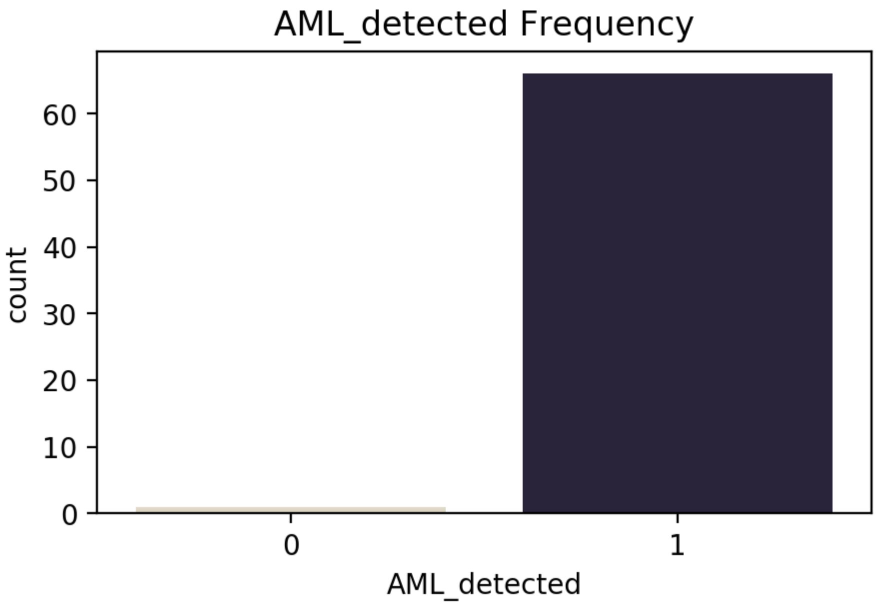


Show 10% Validation Data Split

In [103…
```
clsm_validation.shape
```

Out[103]:
```
(67, 20)
```

In [104…
```
sns.countplot(x=clsm_validation["AML_detected"], palette = "ch:s=-.2,r=.6")
plt.xlabel('AML_detected')
plt.title('AML_detected Frequency')
plt.gcf().set_size_inches(5, 3)
```



In [105…

```
clsm_validation.to_csv('clsm_validation_.csv')
```

In [106…

```
#Manually upload into S3
!aws s3 ls s3://clsm/tabular_regressonehot/
```

```
                           PRE output/
                           PRE test/
                           PRE train/
                           PRE validation/
2023-04-11 01:26:47          0
```

In [107…

```
display(
    HTML(
        '<b>Review <a target="blank" href="https://s3.console.aws.amazon.com/s3/buckets/clsm?prefix=tabular_regressor
            region, account_id, region
        )
    )
)
```

Review S3 Output Bucket

Show 10% Test Data Split

In [108…

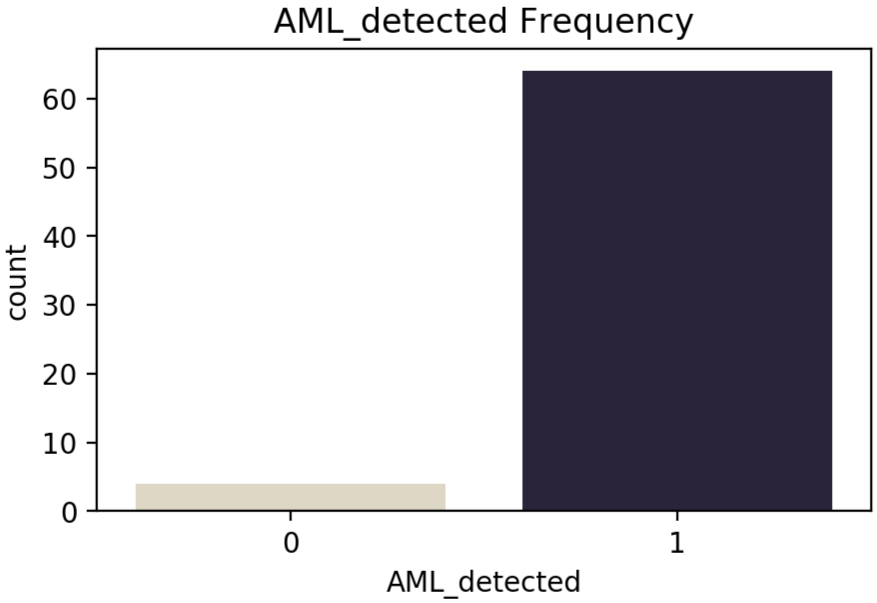
```
clsm_test.shape
```

Out[108]:

```
(68, 20)
```

In [109…

```
sns.countplot(x=clsm_test["AML_detected"], palette = "ch:s=-.2,r=.6")
plt.xlabel('AML_detected')
plt.title('AML_detected Frequency')
plt.gcf().set_size_inches(5, 3)
```



In [110…

```
clsm_test.to_csv('clsm_test_.csv')
```

In [111…

```
#Manually upload into S3
!aws s3 ls s3://clsm/tabular_regressonehot/
```

```
                           PRE output/
                           PRE test/
                           PRE train/
                           PRE validation/
2023-04-11 01:26:47          0
```

In [112...

```
display(
    HTML(
        '<b>Review <a target="blank" href="https://s3.console.aws.amazon.com/s3/buckets/clsm?prefix=tabular_regressor
            region, account_id, region
        )
    )
)
```

Review S3 Output Bucket

Balance Training Dataset

Training Dataset:

In [113...

```
clsm_train["AML_detected"].value_counts()
```

Out[113]:

```
1    516
0     21
Name: AML_detected, dtype: int64
```

Balancing Equation: n = [p(records)-rare]/(1-p), where p=0.50, records=537, rare=21.

In [114...

```
#resampling of training data set
to_resample= clsm_train.loc[clsm_train["AML_detected"] == 0] #isolate all records of AML_detected
our_resample=to_resample.sample(n=495, replace=True) #sample w/ replacement
clsm_t_rebal=pd.concat([clsm_train, our_resample]) #combine original training set w/ resampled records
clsm_t_rebal["AML_detected"].value_counts()
```

Out[114]:

```
1    516
0    516
Name: AML_detected, dtype: int64
```

In [115...

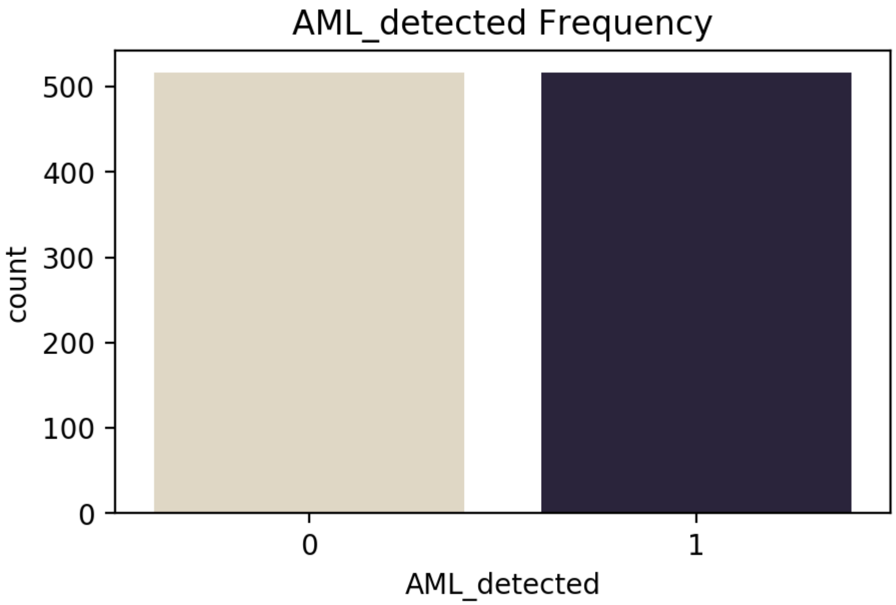
```
clsm_t_rebal.shape
```

Out[115]:

```
(1032, 20)
```

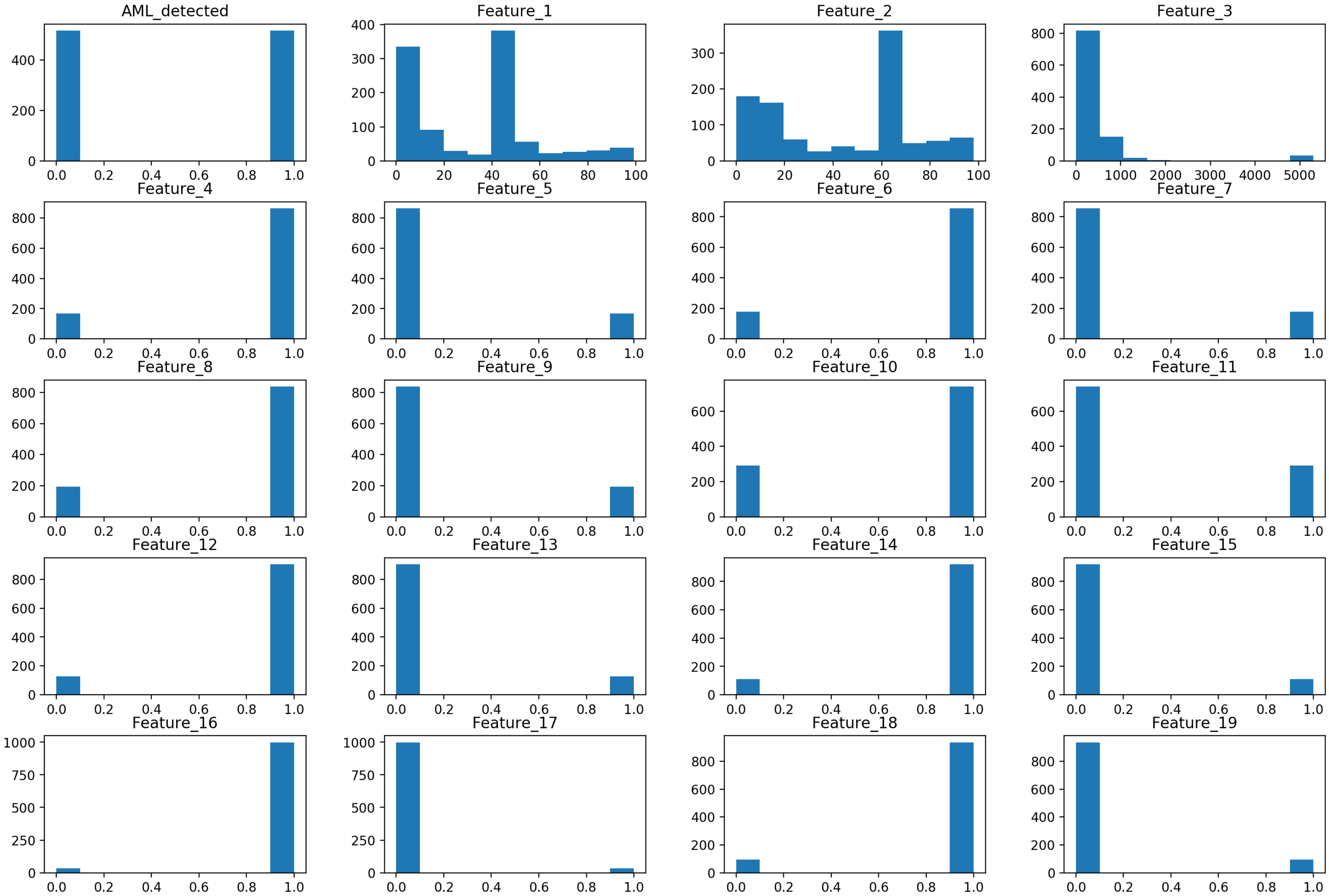
In [116…

```
sns.countplot(x=clsm_t_rebal["AML_detected"], palette = "ch:s=-.2,r=.6")
plt.xlabel('AML_detected')
plt.title('AML_detected Frequency')
plt.gcf().set_size_inches(5, 3)
```



In [117…

```
#clsm_t_rebal Distribution
clsm_t_rebal.hist(grid=False, figsize=(18,12))
plt.show()
```



```
In [118…  clsm_t_rebal.head()
```

Out[118]:

	AML_detected	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5	Feature_6	Feature_7	Feature_8	Feature_9	Feature_10	Featur
480	1	95.0	95.0	201.0	0	1	1	0	1	0	1	
605	1	57.0	40.0	179.0	0	1	1	0	0	1	1	
61	1	41.5	63.0	1993.0	1	0	1	0	1	0	1	
145	1	16.0	63.0	323.0	1	0	1	0	1	0	0	
353	1	0.0	30.0	323.0	1	0	1	0	1	0	1	

In [119…

```
clsm_t_rebal.to_csv('clsm_t_rebal_.csv')
```

In [120…

```
#Manually upload into S3
!aws s3 ls s3://clsm/tabular_regressonehot/
```

```
                           PRE output/
                           PRE test/
                           PRE train/
                           PRE validation/
2023-04-11 01:26:47          0
```

In [121…

```
display(
    HTML(
        '<b>Review <a target="blank" href="https://s3.console.aws.amazon.com/s3/buckets/clsm?prefix=tabular_regressor
            region, account_id, region
        )
    )
)
```

Review S3 Output Bucket

SageMaker JumpStart: XGBoost Model

Set-Up

In [122…

```
!pip install ipywidgets==7.0.0 --quiet
```

```
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system p
ackage manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

[notice] A new release of pip is available: 23.0.1 -> 23.1
[notice] To update, run: pip install --upgrade pip
```

In [123…

```
import json
from sagemaker.session import Session

sagemaker_session = Session()
aws_role = sagemaker_session.get_caller_identity_arn()
aws_region = boto3.Session().region_name
sess = sagemaker.Session()
```

Retrieve Training Artifacts

In [124…

```
model_id, model_version = "xgboost-regression-model", "*"
```

In [125…

```
from ipywidgets import Dropdown
from sagemaker.jumpstart.notebook_utils import list_jumpstart_models
from sagemaker.jumpstart.filters import And

# Retrieves all xgboost and sklearn regression models available by SageMaker Built-In Algorithms.
filter_value = And(f"framework in ['xgboost', 'sklearn']", "task == regression")
text_embedding_models = list_jumpstart_models(filter=filter_value)

# display the model-ids in a dropdown to select a model for inference.
model_dropdown = Dropdown(
    options=text_embedding_models,
    value=model_id,
    description="Select a model",
    style={"description_width": "initial"},
    layout={"width": "max-content"},
)
```

Chose a model for training

In [126…

```
display(model_dropdown)
```

```
A Jupyter Widget
```

In [127…

```
from sagemaker import image_uris, model_uris, script_uris

train_model_id, train_model_version, train_scope = model_dropdown.value, "*", "training"
training_instance_type = "ml.m5.xlarge"

# Retrieve the docker image
train_image_uri = image_uris.retrieve(
    region=None,
    framework=None,
    model_id=train_model_id,
    model_version=train_model_version,
    image_scope=train_scope,
    instance_type=training_instance_type,
)
# Retrieve the training script
train_source_uri = script_uris.retrieve(
    model_id=train_model_id, model_version=train_model_version, script_scope=train_scope
)
# Retrieve the pre-trained model tarball to further fine-tune
train_model_uri = model_uris.retrieve(
    model_id=train_model_id, model_version=train_model_version, model_scope=train_scope
)
```

Set Training Parameters

In [128…

```
# Sample training data is available in this bucket
training_dataset_s3_path = "s3://clsm/tabular_regressonehot/"

s3_output_location = "s3://clsm/tabular_regressonehot/output/"
```

In [129…

```
from sagemaker import hyperparameters

# Retrieve the default hyper-parameters for fine-tuning the model
hyperparameters = hyperparameters.retrieve_default(
    model_id=train_model_id, model_version=train_model_version
)

# [Optional] Override default hyperparameters with custom values
hyperparameters["num_boost_round"] = "500"
hyperparameters["reg_lambda"] = "3"
print(hyperparameters)
```

```
{'num_boost_round': '500', 'early_stopping_rounds': '30', 'learning_rate': '0.3', 'gamma': '0', 'min_child_weight':
'1', 'max_depth': '6', 'subsample': '1', 'colsample_bytree': '1', 'reg_lambda': '3', 'reg_alpha': '0'}
```

Start Training

In [130…

```python
from sagemaker.estimator import Estimator
from sagemaker.utils import name_from_base

training_job_name = name_from_base(f"clsm-rebal-{train_model_id}-training")

# Create SageMaker Estimator instance
tabular_estimator = Estimator(
    role=aws_role,
    image_uri=train_image_uri,
    source_dir=train_source_uri,
    model_uri=train_model_uri,
    entry_point="transfer_learning.py",
    instance_count=1,
    instance_type=training_instance_type,
    max_run=360000,
    hyperparameters=hyperparameters,
    output_path=s3_output_location,
)

# Launch a SageMaker Training job by passing s3 path of the training data
tabular_estimator.fit({"training": "s3://clsm/tabular_regressonehot/train/clsm_t_rebal_.csv",
                       "validation": "s3://clsm/tabular_regressonehot/validation/clsm_validation_.csv" }, logs=True,
```

```
INFO:sagemaker:Creating training-job with name: clsm-rebal-xgboost-regression-model-tra-2023-04-17-02-35-14-214
```

```
2023-04-17 02:35:18 Starting - Starting the training job...
2023-04-17 02:35:34 Starting - Preparing the instances for training...
2023-04-17 02:36:21 Downloading - Downloading input data...
2023-04-17 02:36:43 Training - Downloading the training image...
2023-04-17 02:37:24 Uploading - Uploading generated training model[2023-04-17 02:37:15.789 ip-10-0-175-156.ec2.inter
nal:7 INFO utils.py:28] RULE_JOB_STOP_SIGNAL_FILENAME: None
[2023-04-17 02:37:15.814 ip-10-0-175-156.ec2.internal:7 INFO profiler_config_parser.py:111] User has disabled profil
er.
[2023-04-17:02:37:15:INFO] Imported framework sagemaker_xgboost_container.training
[2023-04-17:02:37:15:INFO] No GPUs detected (normal if no gpus installed)
[2023-04-17:02:37:15:INFO] Invoking user training script.
[2023-04-17:02:37:16:INFO] Module transfer_learning does not provide a setup.py.
Generating setup.py
[2023-04-17:02:37:16:INFO] Generating setup.cfg
[2023-04-17:02:37:16:INFO] Generating MANIFEST.in
[2023-04-17:02:37:16:INFO] Installing module with the following command:
/miniconda3/bin/python3 -m pip install . -r requirements.txt
Processing /opt/ml/code
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Processing ./lib/sagemaker_jumpstart_script_utilities/sagemaker_jumpstart_script_utilities-1.0.1-py2.py3-none-any.wh
l
Building wheels for collected packages: transfer-learning
  Building wheel for transfer-learning (setup.py): started
  Building wheel for transfer-learning (setup.py): finished with status 'done'
  Created wheel for transfer-learning: filename=transfer_learning-1.0.0-py2.py3-none-any.whl size=12553 sha256=ac01a
8e54c7bc5c42d108aaad4296f09a2916e14628573cb9e59a834d6469772
  Stored in directory: /home/model-server/tmp/pip-ephem-wheel-cache-tksdojds/wheels/3e/0f/51/2f1df833dd0412c1bc2f5ee
56baac195b5be563353d111dca6
Successfully built transfer-learning
Installing collected packages: transfer-learning, sagemaker-jumpstart-script-utilities
Successfully installed sagemaker-jumpstart-script-utilities-1.0.1 transfer-learning-1.0.0
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system p
ackage manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
[notice] A new release of pip is available: 23.0.1 -> 23.1
[notice] To update, run: pip install --upgrade pip
[2023-04-17:02:37:17:INFO] No GPUs detected (normal if no gpus installed)
[2023-04-17:02:37:17:INFO] Invoking user script
Training Env:
{
    "additional_framework_parameters": {},
    "channel_input_dirs": {
        "model": "/opt/ml/input/data/model",
        "training": "/opt/ml/input/data/training",
        "validation": "/opt/ml/input/data/validation"
    },
    "current_host": "algo-1",
    "framework_module": "sagemaker_xgboost_container.training:main",
    "hosts": [
        "algo-1"
    ],
    "hyperparameters": {
        "colsample_bytree": "1",
        "early_stopping_rounds": "30",
        "gamma": "0",
        "learning_rate": "0.3",
        "max_depth": "6",
        "min_child_weight": "1",
        "num_boost_round": "500",
        "reg_alpha": "0",
        "reg_lambda": "3",
        "subsample": "1"
    },
    "input_config_dir": "/opt/ml/input/config",
    "input_data_config": {
        "model": {
            "ContentType": "application/x-sagemaker-model",
            "TrainingInputMode": "File",
            "S3DistributionType": "FullyReplicated",
            "RecordWrapperType": "None"
        },
        "training": {
            "TrainingInputMode": "File",
            "S3DistributionType": "FullyReplicated",
            "RecordWrapperType": "None"
        },
        "validation": {
            "TrainingInputMode": "File",
            "S3DistributionType": "FullyReplicated",
            "RecordWrapperType": "None"
        }
    },
    "input_dir": "/opt/ml/input",
    "is_master": true,
    "job_name": "clsm-rebal-xgboost-regression-model-tra-2023-04-17-02-35-14-214",
    "log_level": 20,
    "master_hostname": "algo-1",
    "model_dir": "/opt/ml/model",
    "module_dir": "s3://jumpstart-cache-prod-us-east-1/source-directory-tarballs/xgboost/transfer_learning/regressio
n/v1.1.3/sourcedir.tar.gz",
    "module_name": "transfer_learning",
    "network_interface_name": "eth0",
    "num_cpus": 4,
    "num_gpus": 0,
    "output_data_dir": "/opt/ml/output/data",
    "output_dir": "/opt/ml/output",
    "output_intermediate_dir": "/opt/ml/output/intermediate",
    "resource_config": {
        "current_host": "algo-1",
        "current_instance_type": "ml.m5.xlarge",
        "current_group_name": "homogeneousCluster",
        "hosts": [
            "algo-1"
        ],
        "instance_groups": [
            {
                "instance_group_name": "homogeneousCluster",
                "instance_type": "ml.m5.xlarge",
                "hosts": [
                    "algo-1"
                ]
            }
        ],
        "network_interface_name": "eth0"
    },
    "user_entry_point": "transfer_learning.py"
}
Environment variables:
SM_HOSTS=["algo-1"]
SM_NETWORK_INTERFACE_NAME=eth0
SM_HPS={"colsample_bytree":"1","early_stopping_rounds":"30","gamma":"0","learning_rate":"0.3","max_depth":"6","min_c
hild_weight":"1","num_boost_round":"500","reg_alpha":"0","reg_lambda":"3","subsample":"1"}
SM_USER_ENTRY_POINT=transfer_learning.py
SM_FRAMEWORK_PARAMS={}
SM_RESOURCE_CONFIG={"current_group_name":"homogeneousCluster","current_host":"algo-1","current_instance_type":"ml.m
5.xlarge","hosts":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","inst
ance_type":"ml.m5.xlarge"}],"network_interface_name":"eth0"}
SM_INPUT_DATA_CONFIG={"model":{"ContentType":"application/x-sagemaker-model","RecordWrapperType":"None","S3Distribut
ionType":"FullyReplicated","TrainingInputMode":"File"},"training":{"RecordWrapperType":"None","S3DistributionType":"
FullyReplicated","TrainingInputMode":"File"},"validation":{"RecordWrapperType":"None","S3DistributionType":"FullyRep
licated","TrainingInputMode":"File"}}
SM_OUTPUT_DATA_DIR=/opt/ml/output/data
```

```
SM_CHANNELS=["model","training","validation"]
SM_CURRENT_HOST=algo-1
SM_MODULE_NAME=transfer_learning
SM_LOG_LEVEL=20
SM_FRAMEWORK_MODULE=sagemaker_xgboost_container.training:main
SM_INPUT_DIR=/opt/ml/input
SM_INPUT_CONFIG_DIR=/opt/ml/input/config
SM_OUTPUT_DIR=/opt/ml/output
SM_NUM_CPUS=4
SM_NUM_GPUS=0
SM_MODEL_DIR=/opt/ml/model
SM_MODULE_DIR=s3://jumpstart-cache-prod-us-east-1/source-directory-tarballs/xgboost/transfer_learning/regression/v1.
1.3/sourcedir.tar.gz
SM_TRAINING_ENV={"additional_framework_parameters":{},"channel_input_dirs":{"model":"/opt/ml/input/data/model","trai
ning":"/opt/ml/input/data/training","validation":"/opt/ml/input/data/validation"},"current_host":"algo-1","framework
_module":"sagemaker_xgboost_container.training:main","hosts":["algo-1"],"hyperparameters":{"colsample_bytree":"1","e
arly_stopping_rounds":"30","gamma":"0","learning_rate":"0.3","max_depth":"6","min_child_weight":"1","num_boost_roun
d":"500","reg_alpha":"0","reg_lambda":"3","subsample":"1"},"input_config_dir":"/opt/ml/input/config","input_data_con
fig":{"model":{"ContentType":"application/x-sagemaker-model","RecordWrapperType":"None","S3DistributionType":"FullyR
eplicated","TrainingInputMode":"File"},"training":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicate
d","TrainingInputMode":"File"},"validation":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","Trai
ningInputMode":"File"}},"input_dir":"/opt/ml/input","is_master":true,"job_name":"clsm-rebal-xgboost-regression-model
-tra-2023-04-17-02-35-14-214","log_level":20,"master_hostname":"algo-1","model_dir":"/opt/ml/model","module_dir":"s
3://jumpstart-cache-prod-us-east-1/source-directory-tarballs/xgboost/transfer_learning/regression/v1.1.3/sourcedir.t
ar.gz","module_name":"transfer_learning","network_interface_name":"eth0","num_cpus":4,"num_gpus":0,"output_data_di
r":"/opt/ml/output/data","output_dir":"/opt/ml/output","output_intermediate_dir":"/opt/ml/output/intermediate","reso
urce_config":{"current_group_name":"homogeneousCluster","current_host":"algo-1","current_instance_type":"ml.m5.xlarg
e","hosts":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_ty
pe":"ml.m5.xlarge"}],"network_interface_name":"eth0"},"user_entry_point":"transfer_learning.py"}
SM_USER_ARGS=["--colsample_bytree","1","--early_stopping_rounds","30","--gamma","0","--learning_rate","0.3","--max_d
epth","6","--min_child_weight","1","--num_boost_round","500","--reg_alpha","0","--reg_lambda","3","--subsample","1"]
SM_OUTPUT_INTERMEDIATE_DIR=/opt/ml/output/intermediate
SM_CHANNEL_MODEL=/opt/ml/input/data/model
SM_CHANNEL_TRAINING=/opt/ml/input/data/training
SM_CHANNEL_VALIDATION=/opt/ml/input/data/validation
SM_HP_COLSAMPLE_BYTREE=1
SM_HP_EARLY_STOPPING_ROUNDS=30
SM_HP_GAMMA=0
SM_HP_LEARNING_RATE=0.3
SM_HP_MAX_DEPTH=6
SM_HP_MIN_CHILD_WEIGHT=1
SM_HP_NUM_BOOST_ROUND=500
SM_HP_REG_ALPHA=0
SM_HP_REG_LAMBDA=3
```

```
SM_HP_SUBSAMPLE=1
PYTHONPATH=/miniconda3/bin:/:/miniconda3/lib/python/site-packages/xgboost/dmlc-core/tracker:/miniconda3/lib/python3
7.zip:/miniconda3/lib/python3.7:/miniconda3/lib/python3.7/lib-dynload:/miniconda3/lib/python3.7/site-packages
Invoking script with the following command:
/miniconda3/bin/python3 -m transfer_learning --colsample_bytree 1 --early_stopping_rounds 30 --gamma 0 --learning_ra
te 0.3 --max_depth 6 --min_child_weight 1 --num_boost_round 500 --reg_alpha 0 --reg_lambda 3 --subsample 1
INFO:root:Data in the validation channel is found. Reading the train and validation data from the training and valid
ation channel, respectively.
INFO:root:'_input_model_extracted/__models_info__.json' file could not be found.
[0]#011train-rmse:0.39263#011validation-rmse:0.39462
[1]#011train-rmse:0.29895#011validation-rmse:0.29984
[2]#011train-rmse:0.22434#011validation-rmse:0.21613
[3]#011train-rmse:0.18208#011validation-rmse:0.17288
[4]#011train-rmse:0.14231#011validation-rmse:0.13153
[5]#011train-rmse:0.11810#011validation-rmse:0.10434
[6]#011train-rmse:0.09913#011validation-rmse:0.08728
[7]#011train-rmse:0.08668#011validation-rmse:0.07890
[8]#011train-rmse:0.07221#011validation-rmse:0.06390
[9]#011train-rmse:0.06475#011validation-rmse:0.05671
[10]#011train-rmse:0.05479#011validation-rmse:0.05137
[11]#011train-rmse:0.05005#011validation-rmse:0.04686
[12]#011train-rmse:0.04431#011validation-rmse:0.04442
[13]#011train-rmse:0.04052#011validation-rmse:0.04167
[14]#011train-rmse:0.03802#011validation-rmse:0.03883
[15]#011train-rmse:0.03394#011validation-rmse:0.03641
[16]#011train-rmse:0.03004#011validation-rmse:0.03517
[17]#011train-rmse:0.02525#011validation-rmse:0.03111
[18]#011train-rmse:0.02260#011validation-rmse:0.02985
[19]#011train-rmse:0.02099#011validation-rmse:0.02871
[20]#011train-rmse:0.01903#011validation-rmse:0.02821
[21]#011train-rmse:0.01807#011validation-rmse:0.02742
[22]#011train-rmse:0.01631#011validation-rmse:0.02656
[23]#011train-rmse:0.01576#011validation-rmse:0.02548
[24]#011train-rmse:0.01449#011validation-rmse:0.02505
[25]#011train-rmse:0.01406#011validation-rmse:0.02397
[26]#011train-rmse:0.01316#011validation-rmse:0.02315
[27]#011train-rmse:0.01253#011validation-rmse:0.02258
[28]#011train-rmse:0.01188#011validation-rmse:0.02267
[29]#011train-rmse:0.01150#011validation-rmse:0.02240
[30]#011train-rmse:0.01121#011validation-rmse:0.02218
[31]#011train-rmse:0.01075#011validation-rmse:0.02227
[32]#011train-rmse:0.01045#011validation-rmse:0.02227
[33]#011train-rmse:0.00985#011validation-rmse:0.02199
[34]#011train-rmse:0.00950#011validation-rmse:0.02179
```

```
[35]#011train-rmse:0.00913#011validation-rmse:0.02107
[36]#011train-rmse:0.00875#011validation-rmse:0.02109
[37]#011train-rmse:0.00827#011validation-rmse:0.02083
[38]#011train-rmse:0.00811#011validation-rmse:0.02064
[39]#011train-rmse:0.00779#011validation-rmse:0.02071
[40]#011train-rmse:0.00767#011validation-rmse:0.02060
[41]#011train-rmse:0.00742#011validation-rmse:0.02037
[42]#011train-rmse:0.00714#011validation-rmse:0.02019
[43]#011train-rmse:0.00705#011validation-rmse:0.01991
[44]#011train-rmse:0.00681#011validation-rmse:0.01997
[45]#011train-rmse:0.00655#011validation-rmse:0.01985
[46]#011train-rmse:0.00642#011validation-rmse:0.01969
[47]#011train-rmse:0.00613#011validation-rmse:0.01965
[48]#011train-rmse:0.00603#011validation-rmse:0.01951
[49]#011train-rmse:0.00581#011validation-rmse:0.01979
[50]#011train-rmse:0.00576#011validation-rmse:0.01967
[51]#011train-rmse:0.00567#011validation-rmse:0.01957
[52]#011train-rmse:0.00541#011validation-rmse:0.01951
[53]#011train-rmse:0.00535#011validation-rmse:0.01949
[54]#011train-rmse:0.00509#011validation-rmse:0.01938
[55]#011train-rmse:0.00503#011validation-rmse:0.01936
[56]#011train-rmse:0.00481#011validation-rmse:0.01930
[57]#011train-rmse:0.00470#011validation-rmse:0.01921
[58]#011train-rmse:0.00460#011validation-rmse:0.01912
[59]#011train-rmse:0.00445#011validation-rmse:0.01894
[60]#011train-rmse:0.00430#011validation-rmse:0.01876
[61]#011train-rmse:0.00418#011validation-rmse:0.01866
[62]#011train-rmse:0.00408#011validation-rmse:0.01858
[63]#011train-rmse:0.00393#011validation-rmse:0.01840
[64]#011train-rmse:0.00385#011validation-rmse:0.01843
[65]#011train-rmse:0.00375#011validation-rmse:0.01837
[66]#011train-rmse:0.00364#011validation-rmse:0.01831
[67]#011train-rmse:0.00352#011validation-rmse:0.01831
[68]#011train-rmse:0.00345#011validation-rmse:0.01827
[69]#011train-rmse:0.00339#011validation-rmse:0.01827
[70]#011train-rmse:0.00332#011validation-rmse:0.01826
[71]#011train-rmse:0.00328#011validation-rmse:0.01825
[72]#011train-rmse:0.00324#011validation-rmse:0.01824
[73]#011train-rmse:0.00321#011validation-rmse:0.01824
[74]#011train-rmse:0.00314#011validation-rmse:0.01822
[75]#011train-rmse:0.00311#011validation-rmse:0.01823
[76]#011train-rmse:0.00301#011validation-rmse:0.01824
[77]#011train-rmse:0.00295#011validation-rmse:0.01818
[78]#011train-rmse:0.00286#011validation-rmse:0.01827
```

```
[79]#011train-rmse:0.00281#011validation-rmse:0.01822
[80]#011train-rmse:0.00278#011validation-rmse:0.01816
[81]#011train-rmse:0.00274#011validation-rmse:0.01813
[82]#011train-rmse:0.00269#011validation-rmse:0.01808
[83]#011train-rmse:0.00266#011validation-rmse:0.01806
[84]#011train-rmse:0.00264#011validation-rmse:0.01802
[85]#011train-rmse:0.00259#011validation-rmse:0.01803
[86]#011train-rmse:0.00255#011validation-rmse:0.01795
[87]#011train-rmse:0.00252#011validation-rmse:0.01798
[88]#011train-rmse:0.00246#011validation-rmse:0.01803
[89]#011train-rmse:0.00243#011validation-rmse:0.01802
[90]#011train-rmse:0.00238#011validation-rmse:0.01801
[91]#011train-rmse:0.00237#011validation-rmse:0.01801
[92]#011train-rmse:0.00231#011validation-rmse:0.01809
[93]#011train-rmse:0.00226#011validation-rmse:0.01806
[94]#011train-rmse:0.00220#011validation-rmse:0.01811
[95]#011train-rmse:0.00214#011validation-rmse:0.01816
[96]#011train-rmse:0.00208#011validation-rmse:0.01815
[97]#011train-rmse:0.00198#011validation-rmse:0.01797
[98]#011train-rmse:0.00197#011validation-rmse:0.01796
[99]#011train-rmse:0.00192#011validation-rmse:0.01796
[100]#011train-rmse:0.00190#011validation-rmse:0.01796
[101]#011train-rmse:0.00189#011validation-rmse:0.01795
[102]#011train-rmse:0.00187#011validation-rmse:0.01794
[103]#011train-rmse:0.00187#011validation-rmse:0.01794
[104]#011train-rmse:0.00187#011validation-rmse:0.01794
[105]#011train-rmse:0.00187#011validation-rmse:0.01794
[106]#011train-rmse:0.00187#011validation-rmse:0.01794
[107]#011train-rmse:0.00187#011validation-rmse:0.01794
[108]#011train-rmse:0.00187#011validation-rmse:0.01794
[109]#011train-rmse:0.00187#011validation-rmse:0.01794
[110]#011train-rmse:0.00187#011validation-rmse:0.01794
[111]#011train-rmse:0.00187#011validation-rmse:0.01794
[112]#011train-rmse:0.00187#011validation-rmse:0.01794
[113]#011train-rmse:0.00187#011validation-rmse:0.01794
[114]#011train-rmse:0.00187#011validation-rmse:0.01794
[115]#011train-rmse:0.00187#011validation-rmse:0.01794
[116]#011train-rmse:0.00187#011validation-rmse:0.01794
[117]#011train-rmse:0.00187#011validation-rmse:0.01794
[118]#011train-rmse:0.00187#011validation-rmse:0.01794
[119]#011train-rmse:0.00187#011validation-rmse:0.01794
[120]#011train-rmse:0.00187#011validation-rmse:0.01794
[121]#011train-rmse:0.00187#011validation-rmse:0.01794
[122]#011train-rmse:0.00187#011validation-rmse:0.01794
```

```
[123]#011train-rmse:0.00187#011validation-rmse:0.01794
[124]#011train-rmse:0.00187#011validation-rmse:0.01794
[125]#011train-rmse:0.00187#011validation-rmse:0.01794
[126]#011train-rmse:0.00187#011validation-rmse:0.01794
[127]#011train-rmse:0.00187#011validation-rmse:0.01794
[128]#011train-rmse:0.00187#011validation-rmse:0.01794
[129]#011train-rmse:0.00187#011validation-rmse:0.01794
[130]#011train-rmse:0.00187#011validation-rmse:0.01794
[131]#011train-rmse:0.00187#011validation-rmse:0.01794
[132]#011train-rmse:0.00187#011validation-rmse:0.01794
INFO:root:Saving model...
INFO:root:Info file not found at '_input_model_extracted/__models_info__.json'.

2023-04-17 02:37:36 Completed - Training job completed
Training seconds: 78
Billable seconds: 78
```

Deploy and Run Inference on the Trained Tabular Model

In [131…

```python
inference_instance_type = "ml.m5.large"

# Retrieve the inference docker container uri
deploy_image_uri = image_uris.retrieve(
    region=None,
    framework=None,
    image_scope="inference",
    model_id=train_model_id,
    model_version=train_model_version,
    instance_type=inference_instance_type,
)
# Retrieve the inference script uri
deploy_source_uri = script_uris.retrieve(
    model_id=train_model_id, model_version=train_model_version, script_scope="inference"
)

endpoint_name = name_from_base(f"clsm-train-{train_model_id}-")

# Use the estimator from the previous step to deploy to a SageMaker endpoint
predictor = tabular_estimator.deploy(
    initial_instance_count=1,
    instance_type=inference_instance_type,
    entry_point="inference.py",
    image_uri=deploy_image_uri,
    source_dir=deploy_source_uri,
    endpoint_name=endpoint_name,
    enable_network_isolation=True,
)
```

```
INFO:sagemaker.image_uris:Ignoring unnecessary Python version: py3.
INFO:sagemaker.image_uris:Ignoring unnecessary instance type: ml.m5.large.
INFO:sagemaker:Creating model with name: sagemaker-jumpstart-2023-04-17-02-38-07-750
INFO:sagemaker:Creating endpoint-config with name clsm-train-xgboost-regression-model--2023-04-17-02-38-07-750
INFO:sagemaker:Creating endpoint with name clsm-train-xgboost-regression-model--2023-04-17-02-38-07-750
----!
```

Test Data

In [132…

```python
newline, bold, unbold = "\n", "\033[1m", "\033[0m"

from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt

# read the data
test_data = clsm_test
test_data.columns = ["AML_detected"] + [f"Feature_{i}" for i in range(1, test_data.shape[1])]
num_examples, num_columns = test_data.shape
print(
    f"{bold}The test dataset contains {num_examples} examples and {num_columns} columns.{unbold}\n"
)

# prepare the ground truth target and predicting features to send into the endpoint.
ground_truth_label, features = test_data.iloc[:, :1] , test_data.iloc[:, 1:]

print(
    f"{bold}The first 5 observations of the test data: {unbold}"
)  # Feature_1 is the categorical variables and rest of other features are numeric variables.
test_data.head(5)
```

```
The test dataset contains 68 examples and 20 columns.

The first 5 observations of the test data:
```

Out[132]:

	AML_detected	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5	Feature_6	Feature_7	Feature_8	Feature_9	Feature_10	Featur
318	1	55.0	51.0	35.0	0	1	1	0	1	0	1	
300	1	22.0	36.0	854.0	1	0	1	0	1	0	1	
478	1	26.7	25.0	637.0	1	0	1	0	0	1	1	
155	1	6.0	10.0	357.0	1	0	1	0	1	0	1	
31	1	75.0	77.5	1735.0	0	1	0	1	1	0	1	

Predict

In [133…

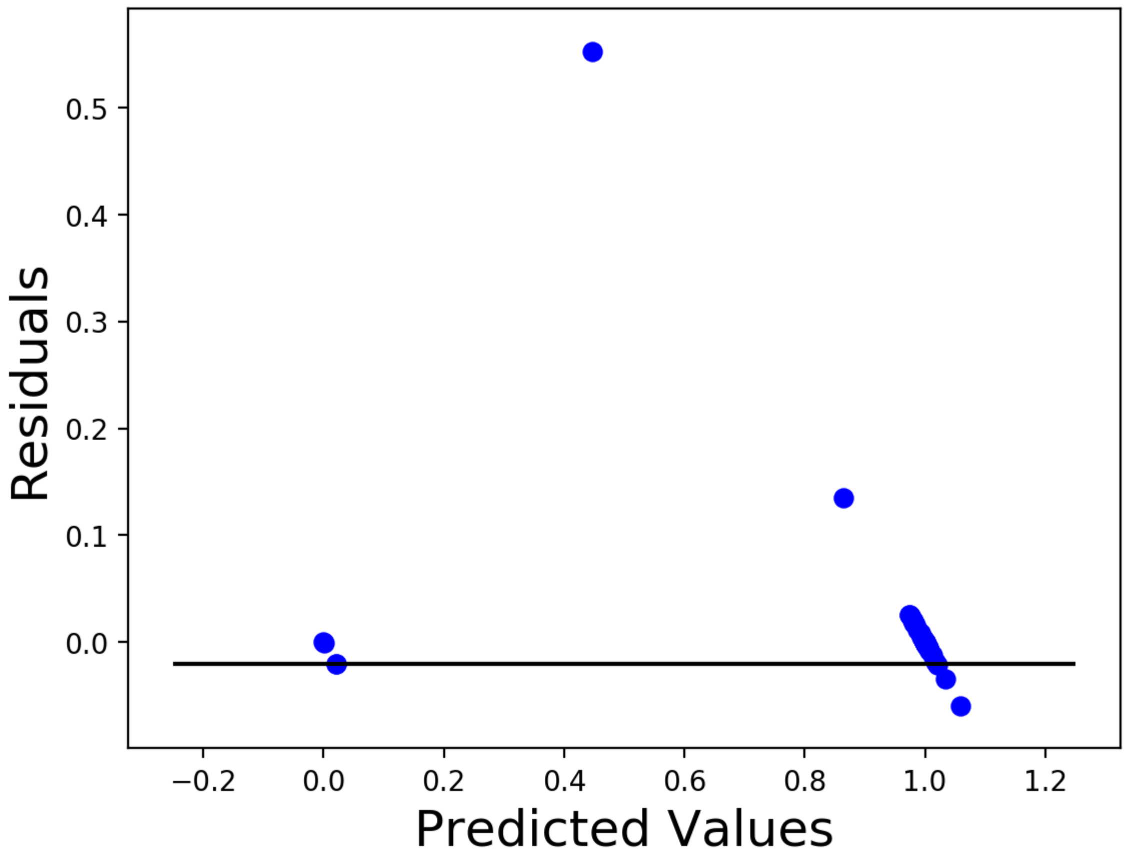
```python
content_type = "text/csv"


def query_endpoint(encoded_tabular_data):
    client = boto3.client("runtime.sagemaker")
    response = client.invoke_endpoint(
        EndpointName=endpoint_name, ContentType=content_type, Body=encoded_tabular_data
    )
    return response


def parse_resonse(query_response):
    predictions = json.loads(query_response["Body"].read())
    return np.array(predictions["prediction"])


query_response = query_endpoint(features.to_csv(header=False, index=False).encode("utf-8"))
model_predictions = parse_resonse(query_response)
```

Evaluate Predictions

Visual

In [134…

```python
# Visualization: a residual plot to compare the model predictions and ground truth targets.
# Binary results

residuals = ground_truth_label.values[:, 0] - model_predictions
plt.scatter(model_predictions, residuals, color="blue", s=40)
plt.hlines(y=-0.02, xmin=-0.25, xmax=1.25)
plt.xlabel("Predicted Values", fontsize=18)
plt.ylabel("Residuals", fontsize=18)
plt.show()
```



Quantitative

In [135…

```
# Evaluate the model predictions quantitatively.
eval_r2_score = r2_score(ground_truth_label.values, model_predictions)
eval_mse_score = mean_squared_error(ground_truth_label.values, model_predictions)
eval_mae_score = mean_absolute_error(ground_truth_label.values, model_predictions)
print(
    f"{bold}Evaluation result on test data{unbold}:{newline}"
    f"{bold}{r2_score.__name__}{unbold}: {eval_r2_score}{newline}"
    f"{bold}{mean_squared_error.__name__}{unbold}: {eval_mse_score}{newline}"
    f"{bold}{mean_absolute_error.__name__}{unbold}: {eval_mae_score}{newline}"
)
```

```
Evaluation result on test data:
r2_score: 0.9111131505823139
mean_squared_error: 0.0049210712480379825
mean_absolute_error: 0.01730402212791994
```

Delete SageMaker Endpoint

In [136…

```
# Delete the SageMaker endpoint and the attached resources
predictor.delete_model()
predictor.delete_endpoint()
```

```
INFO:sagemaker:Deleting model with name: sagemaker-jumpstart-2023-04-17-02-38-07-750
INFO:sagemaker:Deleting endpoint configuration with name: clsm-train-xgboost-regression-model--2023-04-17-02-38-07-7
50
INFO:sagemaker:Deleting endpoint with name: clsm-train-xgboost-regression-model--2023-04-17-02-38-07-750
```

XGBoost Metrics

In [137…

```
import sklearn.metrics as metrics
from sklearn.metrics import classification_report, roc_curve
```

In [138…

```
# Target values = Test dataset AML Detected

target = ground_truth_label.values[:, 0]
print(target)
```

```
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1]
```

In [139…

```
# Test dataset Predicted values

pred = (model_predictions> 0.5).astype(np.float32)
print(pred)
```

```
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1.]
```

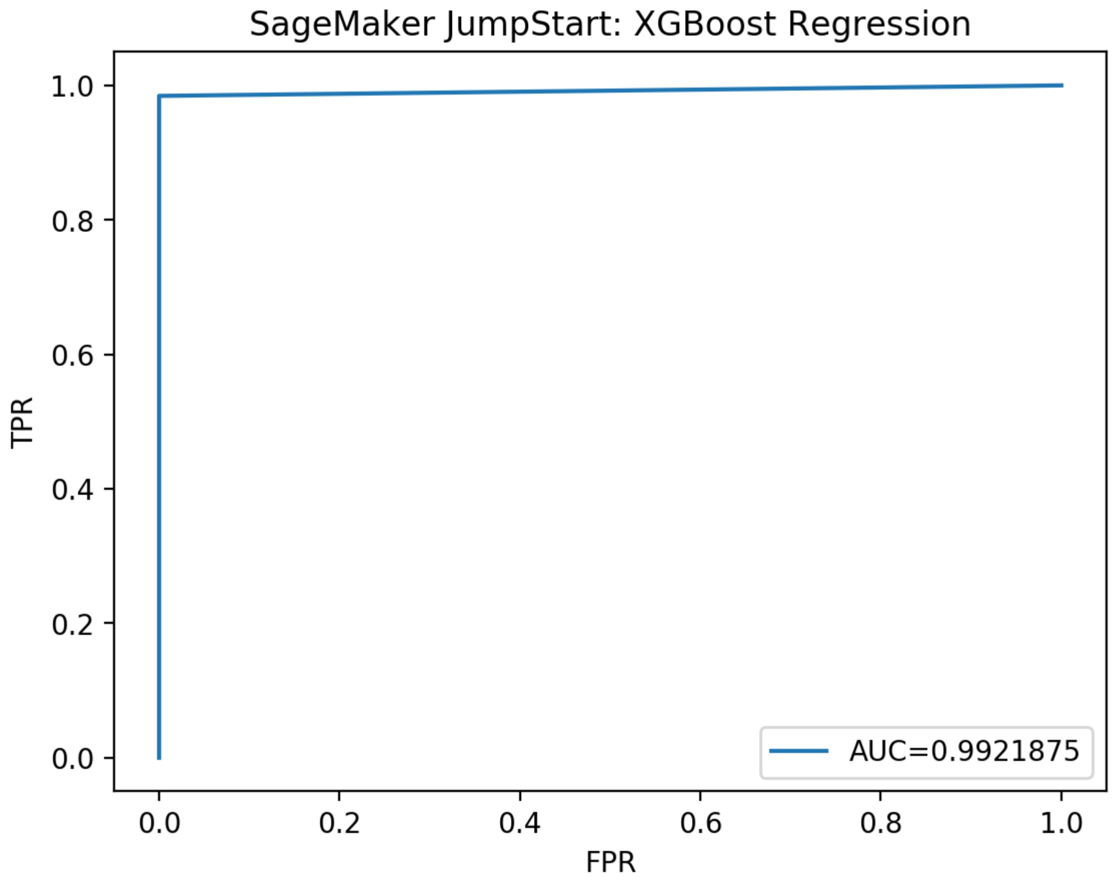
In [140…

```
#Cross Validation

ALM_detected = ['no', 'yes']
print('Cross Validation: \n',
      classification_report(target, pred, target_names=ALM_detected))
```

```
Cross Validation:
               precision    recall  f1-score   support

          no       0.80      1.00      0.89         4
         yes       1.00      0.98      0.99        64

    accuracy                           0.99        68
   macro avg       0.90      0.99      0.94        68
weighted avg       0.99      0.99      0.99        68
```

In [141…

```
#ROC curve for SageMaker JumpStart: XGBoost Regression
fpr, tpr, _ = metrics.roc_curve(target,  pred)
auc = metrics.roc_auc_score(target, pred)

plt.plot(fpr,tpr,label="AUC="+str(auc))
plt.legend(loc=4)
plt.title('SageMaker JumpStart: XGBoost Regression')
plt.xlabel('FPR')
plt.ylabel('TPR')
plt.show()
```



Release Resources

In [142…

```
%%html

<p><b>Shutting down your kernel for this notebook to release resources.</b></p>
<button class="sm-command-button" data-commandlinker-command="kernelmenu:shutdown" style="display:none;">Shutdown Ker

<script>
try {
    els = document.getElementsByClassName("sm-command-button");
    els[0].click();
}
catch(err) {
    // NoOp
}
</script>
```

Shutting down your kernel for this notebook to release resources.

In [143…

```
%%javascript

try {
    Jupyter.notebook.save_checkpoint();
    Jupyter.notebook.session.delete();
}
catch(err) {
    // NoOp
}
```

References

AWS SageMaker. Jumpstart. SGBoost Regression Model. SageMaker Built-In Algorithms: Tabular Regression using XGBoost and Linear Learner

In [ ]:

```
```