GoogleCloudPlatform / **mlops-on-gcp** Public

Code Issues 23 Pull requests 8 Actions Projects Security

master

mlops-on-gcp / datasets / covertype / wrangle / **prepare.ipynb**

jarokaz Initial draft History

0 contributors

2414 lines (2414 sloc) | 83.4 KB

Covertype Data Set Preprocessing

InÂ [6]:

```
import csv
import numpy as np
import pandas as pd

from sklearn.model_selection import train_test_split
```

Set the paths

InÂ [7]:

```
FULL_DATASET = '../covertype.csv'
SMALL_DATASET= '../covertype_small.csv'
TRAINING_DATASET='../covertype_training.csv'
TRAINING_DATASET_WITH_MISSING = '../covertype_training_missing.csv'
EVALUATION_DATASET='../covertype_evaluation.csv'
EVALUATION_DATASET_WITH_ANOMALIES='../covertype_evaluation_anomalie
SERVING_DATASET='../covertype_serving.csv'

ORIGINAL_DATASET_PATH = 'gs://workshop-datasets/covertype/orig/covt
```

Preprocess the original dataset

Load the dataset

InÂ [8]:

```
df = pd.read_csv(ORIGINAL_DATASET_PATH, header=None)
print(df.shape)
df.head()
```

```
(581012, 55)
```

Out[8]:

	0	1	2	3	4	5	6	7	8	9	...	45	46	47	48	49	50	5
0	2596	51	3	258	0	510	221	232	148	6279	...	0	0	0	0	0	0	
1	2590	56	2	212	-6	390	220	235	151	6225	...	0	0	0	0	0	0	
2	2804	139	9	268	65	3180	234	238	135	6121	...	0	0	0	0	0	0	
3	2785	155	18	242	118	3090	238	238	122	6211	...	0	0	0	0	0	0	
4	2595	45	2	153	-1	391	220	234	150	6172	...	0	0	0	0	0	0	

5 rows ÃƒÂ— 55 columns

Configure soil type and wilderness area domains

InÂ [10…

```
soil_type = [
"1", "C2702", "Cathedral family - Rock outcrop complex, extremely s
"2", "C2703", "Vanet - Ratake families complex, very stony.",
"3", "C2704", "Haploborolis - Rock outcrop complex, rubbly.",
"4", "C2705", "Ratake family - Rock outcrop complex, rubbly.",
"5", "C2706", "Vanet family - Rock outcrop complex complex, rubbly.
```

```
"6", "C2717", "Vanet - Wetmore families - Rock outcrop complex, sto
"7", "C3501", "Gothic family.",
"8", "C3502", "Supervisor - Limber families complex.",
"9", "C4201", "Troutville family, very stony.",
"10", "C4703", "Bullwark - Catamount families - Rock outcrop comple
"11", "C4704", "Bullwark - Catamount families - Rock land complex,
"12", "C4744", "Legault family - Rock land complex, stony.",
"13", "C4758", "Catamount family - Rock land - Bullwark family comp
"14", "C5101", "Pachic Argiborolis - Aquolis complex.",
"15", "C5151", "unspecified in the USFS Soil and ELU Survey.",
"16", "C6101", "Cryaquolis - Cryoborolis complex.",
"17", "C6102", "Gateview family - Cryaquolis complex.",
"18", "C6731", "Rogert family, very stony.",
"19", "C7101", "Typic Cryaquolis - Borohemists complex.",
"20", "C7102", "Typic Cryaquepts - Typic Cryaquolls complex.",
"21", "C7103", "Typic Cryaquolls - Leighcan family, till substratum
"22", "C7201", "Leighcan family, till substratum, extremely boulder
"23", "C7202", "Leighcan family, till substratum - Typic Cryaquolls
"24", "C7700", "Leighcan family, extremely stony.",
"25", "C7701", "Leighcan family, warm, extremely stony.",
"26", "C7702", "Granile - Catamount families complex, very stony.",
"27", "C7709", "Leighcan family, warm - Rock outcrop complex, extre
"28", "C7710", "Leighcan family - Rock outcrop complex, extremely s
"29", "C7745", "Como - Legault families complex, extremely stony.",
"30", "C7746", "Como family - Rock land - Legault family complex, e
"31", "C7755", "Leighcan - Catamount families complex, extremely st
"32", "C7756", "Catamount family - Rock outcrop - Leighcan family c
"33", "C7757", "Leighcan - Catamount families - Rock outcrop comple
"34", "C7790", "Cryorthents - Rock land complex, extremely stony.",
"35", "C8703", "Cryumbrepts - Rock outcrop - Cryaquepts complex.",
"36", "C8707", "Bross family - Rock land - Cryumbrepts complex, ext
"37", "C8708", "Rock outcrop - Cryumbrepts - Cryorthents complex, e
"38", "C8771", "Leighcan - Moran families - Cryaquolls complex, ext
"39", "C8772", "Moran family - Cryorthents - Leighcan family comple
"40", "C8776", "Moran family - Cryorthents - Rock land complex, ext
]

wilderness_area = [
"Rawah", "Rawah Wilderness Area",
"Neota", "Neota Wilderness Area",
"Commanche", "Comanche Peak Wilderness Area",
"Cache", "Cache la Poudre Wilderness Area"
]
```

Map one-hot encoded values to categorical domains

InÂ [11…

```
soil = df.loc[:, 14:53].apply(lambda x: soil_type[1::3][x.to_numpy(
soil
```

Out[11]:

```
0         C7745
1         C7745
2         C4744
3         C7746
4         C7745
          ...
581007    C2703
581008    C2703
581009    C2703
581010    C2703
581011    C2703
Length: 581012, dtype: object
```

```
Length: 581012, dtype: object
```

InÂ [12…

```
wilderness = df.loc[:, 10:13].apply(lambda x: wilderness_area[0::2]
wilderness
```

Out[12]:

```
0             Rawah
1             Rawah
2             Rawah
3             Rawah
4             Rawah
            ...
581007    Commanche
581008    Commanche
581009    Commanche
581010    Commanche
581011    Commanche
Length: 581012, dtype: object
```

Create a dataset with column names and categorical values replacing one-hot encoded soil type and wilderness areas

InÂ [13…

```
COLUMN_NAMES = [
    'Elevation',
    'Aspect',
    'Slope',
    'Horizontal_Distance_To_Hydrology',
    'Vertical_Distance_To_Hydrology',
    'Horizontal_Distance_To_Roadways',
    'Hillshade_9am',
    'Hillshade_Noon',
    'Hillshade_3pm',
    'Horizontal_Distance_To_Fire_Points',
    'Wilderness_Area',
    'Soil_Type',
    'Cover_Type']

df_full = pd.concat([df.loc[:, 0:9], wilderness, soil, df.loc[:, 54
df_full.columns = COLUMN_NAMES
df_full
```

Out[13]:

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_
0	2596	51	3	258	
1	2590	56	2	212	
2	2804	139	9	268	
3	2785	155	18	242	
4	2595	45	2	153	
...	...	...	...	...	
581007	2396	153	20	85	
581008	2391	152	19	67	
581009	2386	159	17	60	
581010	2384	170	15	60	
581011	2383	165	13	60	

581012 rows ÃƒÂ— 13 columns

Convert the label to 0-6 range

InÂ [15…

```
df_full['Cover_Type'] = df_full['Cover_Type'] - 1
```

Save the dataset to CSV file

InÂ [18…

```
df_full.to_csv(FULL_DATASET, header=True, index=False)
```

InÂ [19…

```
!head $FULL_DATASET
```

```
Elevation,Aspect,Slope,Horizontal_Distance_To_Hydrology,Vertical_Di
stance_To_Hydrology,Horizontal_Distance_To_Roadways,Hillshade_9am,H
illshade_Noon,Hillshade_3pm,Horizontal_Distance_To_Fire_Points,Wild
erness_Area,Soil_Type,Cover_Type
2596,51,3,258,0,510,221,232,148,6279,Rawah,C7745,4
2590,56,2,212,-6,390,220,235,151,6225,Rawah,C7745,4
2804,139,9,268,65,3180,234,238,135,6121,Rawah,C4744,1
2785,155,18,242,118,3090,238,238,122,6211,Rawah,C7746,1
2595,45,2,153,-1,391,220,234,150,6172,Rawah,C7745,4
2579,132,6,300,-15,67,230,237,140,6031,Rawah,C7745,1
2606,45,7,270,5,633,222,225,138,6256,Rawah,C7745,4
2605,49,4,234,7,573,222,230,144,6228,Rawah,C7745,4
2617,45,9,240,56,666,223,221,133,6244,Rawah,C7745,4
```

Create training, validation, testing and serving splits.

InÂ [20…

```
df_full = df = pd.read_csv(FULL_DATASET, dtype={'Soil_Type': object
df_full
```

Out[20]:

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_
0	2596	51	3	258	
1	2590	56	2	212	
2	2804	139	9	268	
3	2785	155	18	242	
4	2595	45	2	153	
...	...	...	...	...	
581007	2396	153	20	85	
581008	2391	152	19	67	
581009	2386	159	17	60	
581010	2384	170	15	60	
581011	2383	165	13	60	

581012 rows ÃƒÂ— 13 columns

InÂ [21…

```
df_full.Soil_Type.value_counts()
```

Out[21]:

```
C7745    115247
C7202     57752
C7756     52519
C7757     45154
C7201     33373
C4703     32634
C7746     30170
C4744     29971
C7755     25666
C7700     21278
C4758     17431
C8771     15573
C8772     13806
C4704     12410
C2705     12396
C7102      9259
C8776      8750
C2703      7525
C2717      6575
C2704      4823
C7101      4021
C6102      3422
C2702      3031
C6101      2845
C7702      2589
C6731      1899
C8703      1891
C7790      1611
C2706      1597
C4201      1147
C7709      1086
C7710       946
C7103       838
C5101       599
C7701       474
C8708       298
C3502       179
C8707       119
C3501       105
C5151         3
Name: Soil_Type, dtype: int64
```

InÂ [22…

```
df_5151 = df_full[df_full['Soil_Type']=='C5151']
df_no_5151 = df_full[df_full['Soil_Type']!='C5151']
```

InÂ [23…

```
df_5151
```

Out[23]:

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_
241543	2078	34	10	0	
241544	2080	13	19	30	
241545	2076	27	24	30	

InÂ [24…

```
df_no_5151
```

Out[24]:

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_
0	2596	51	3	258	
1	2590	56	2	212	
2	2804	139	9	268	
3	2785	155	18	242	
4	2595	45	2	153	
...	...	...	...	...	
581007	2396	153	20	85	
581008	2391	152	19	67	
581009	2386	159	17	60	
581010	2384	170	15	60	
581011	2383	165	13	60	

581009 rows ÃƒÂ— 13 columns

InÂ [25…

```
df_small, df_other = train_test_split(df_no_5151, train_size=100000
```

InÂ [26…

```
df_train, df_other = train_test_split(df_no_5151, train_size=431009
df_evaluate, df_serving = train_test_split(df_other, train_size=750
df_serving = df_serving.drop(columns=['Cover_Type'])
print(df_train.shape)
print(df_evaluate.shape)
print(df_serving.shape)
```

```
(431009, 13)
(75000, 13)
(75000, 12)
```

Add some missing values to the training split.

InÂ [27…

```
df_train_missing = df_train.reset_index(drop=True)
df_train_missing.loc[0:8999, 'Horizontal_Distance_To_Hydrology'] =
df_train_missing
```

Out[27]:

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_
0	3212	191	10	NaN	
1	3205	3	14	NaN	
2	2605	74	16	NaN	
3	2768	73	31	NaN	
4	3230	45	13	NaN	
...	...	...	...	...	
431004	2950	36	4	108.0	
431005	2837	278	10	30.0	
431006	3101	152	9	150.0	

431007	3228	136	14	216.0
431008	3060	358	16	495.0

431009 rows ÃƒÂ— 13 columns

Create the evaluation split where some values of Slope are more than 90 degrees and 3 examples have 5151 code for soil type, which is not present in the training split.

InÂ [28…

```
df_evaluate_anomalies = df_evaluate.reset_index(drop=True)
df_evaluate_anomalies.loc[0:4, 'Slope'] = 110
df_evaluate_anomalies = pd.concat([df_evaluate_anomalies, df_5151])
df_evaluate_anomalies
```

Out[28]:

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_
0	3001	96	110	534	
1	3005	139	110	175	
2	2768	91	110	242	
3	3153	346	110	277	
4	3379	68	110	150	
...	...	...	...	...	
74998	2958	61	24	234	
74999	3159	132	14	150	
241543	2078	34	10	0	
241544	2080	13	19	30	
241545	2076	27	24	30	

75003 rows ÃƒÂ— 13 columns

InÂ [29…

```
df_evaluate_anomalies.Soil_Type.value_counts()
```

Out[29]:

```
C7745    14939
C7202     7512
C7756     6731
C7757     5740
C7201     4249
C4703     4167
C4744     3905
C7746     3851
C7755     3408
C7700     2838
C4758     2231
C8771     1992
C8772     1841
C4704     1598
C2705     1592
C7102     1091
C8776     1078
C2703      968
```

```
C2717      871
C2704      636
C7101      489
C6102      452
C2702      420
C6101      383
C7702      316
C8703      257
C6731      242
C7790      214
C2706      211
C4201      147
C7709      145
C7710      135
C7103      109
C5101       74
C7701       63
C8708       52
C3502       21
C3501       16
C8707       16
C5151        3
Name: Soil_Type, dtype: int64
```

Save the splits to local files.

InÂ [30…

```
df_train.to_csv(TRAINING_DATASET, header=True, index=False)
df_small.to_csv(SMALL_DATASET, header=True, index=False)
df_train_missing.to_csv(TRAINING_DATASET_WITH_MISSING, header=True,
df_evaluate.to_csv(EVALUATION_DATASET, header=True, index=False)
df_evaluate_anomalies.to_csv(EVALUATION_DATASET_WITH_ANOMALIES, hea
df_serving.to_csv(SERVING_DATASET, header=True, index=False)
```

Copy the splits to GCS

InÂ [31…

```
!gsutil cp $FULL_DATASET gs://workshop-datasets/covertype/full/data
!gsutil cp $SMALL_DATASET gs://workshop-datasets/covertype/small/da
!gsutil cp $TRAINING_DATASET gs://workshop-datasets/covertype/train
!gsutil cp $TRAINING_DATASET_WITH_MISSING gs://workshop-datasets/co
!gsutil cp $EVALUATION_DATASET gs://workshop-datasets/covertype/eva
!gsutil cp $EVALUATION_DATASET_WITH_ANOMALIES gs://workshop-dataset
!gsutil cp $SERVING_DATASET gs://workshop-datasets/covertype/servin
```

```
Copying file://../covertype.csv [Content-Type=text/csv]...
- [1 files][ 30.5 MiB/ 30.5 MiB]
Operation completed over 1 objects/30.5 MiB.
Copying file://../covertype_small.csv [Content-Type=text/csv]...
/ [1 files][  5.3 MiB/  5.3 MiB]
Operation completed over 1 objects/5.3 MiB.
Copying file://../covertype_training.csv [Content-Type=text/csv]...
- [1 files][ 22.7 MiB/ 22.7 MiB]
Operation completed over 1 objects/22.7 MiB.
Copying file://../covertype_training_missing.csv [Content-Type=text
/csv]...
- [1 files][ 23.4 MiB/ 23.4 MiB]
Operation completed over 1 objects/23.4 MiB.
Copying file://../covertype_evaluation.csv [Content-Type=text/cs
v]...
/ [1 files][  3.9 MiB/  3.9 MiB]
```

```
Operation completed over 1 objects/3.9 MiB.
```