Feature Engineering

In this notebook, you will learn how to incorporate feature engineering into your pipeline.

- Working with feature columns
- Adding feature crosses in TensorFlow
- Reading data from BigQuery
- Creating datasets using Dataflow
- Using a wide-and-deep model

In [1]:
```
!sudo chown -R jupyter:jupyter /home/jupyter/training-data-analyst
```

In [2]:
```
!pip install --user google-cloud-bigquery==1.25.0
```

```
Collecting google-cloud-bigquery==1.25.0
  Downloading google_cloud_bigquery-1.25.0-py2.py3-none-any.whl (169 kB)
     |████████████████████████████████| 169 kB 20.7 MB/s eta 0:00:01
Requirement already satisfied: protobuf>=3.6.0 in /opt/conda/lib/python3.7/site-packages
(from google-cloud-bigquery==1.25.0) (3.16.0)
Requirement already satisfied: six<2.0.0dev,>=1.13.0 in /opt/conda/lib/python3.7/site-pa
ckages (from google-cloud-bigquery==1.25.0) (1.16.0)
Requirement already satisfied: google-api-core<2.0dev,>=1.15.0 in /opt/conda/lib/python
3.7/site-packages (from google-cloud-bigquery==1.25.0) (1.26.3)
Requirement already satisfied: google-auth<2.0dev,>=1.9.0 in /opt/conda/lib/python3.7/si
te-packages (from google-cloud-bigquery==1.25.0) (1.30.0)
Collecting google-resumable-media<0.6dev,>=0.5.0
  Downloading google_resumable_media-0.5.1-py2.py3-none-any.whl (38 kB)
Requirement already satisfied: google-cloud-core<2.0dev,>=1.1.0 in /opt/conda/lib/python
3.7/site-packages (from google-cloud-bigquery==1.25.0) (1.6.0)
Requirement already satisfied: packaging>=14.3 in /opt/conda/lib/python3.7/site-packages
(from google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==1.25.0) (20.9)
Requirement already satisfied: setuptools>=40.3.0 in /opt/conda/lib/python3.7/site-packa
ges (from google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==1.25.0) (49.6.0.post20
210108)
Requirement already satisfied: requests<3.0.0dev,>=2.18.0 in /opt/conda/lib/python3.7/si
te-packages (from google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==1.25.0) (2.25.
1)
Requirement already satisfied: pytz in /opt/conda/lib/python3.7/site-packages (from goog
le-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==1.25.0) (2021.1)
Requirement already satisfied: googleapis-common-protos<2.0dev,>=1.6.0 in /opt/conda/li
b/python3.7/site-packages (from google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==
1.25.0) (1.53.0)
Requirement already satisfied: pyasn1-modules>=0.2.1 in /opt/conda/lib/python3.7/site-pa
ckages (from google-auth<2.0dev,>=1.9.0->google-cloud-bigquery==1.25.0) (0.2.7)
Requirement already satisfied: cachetools<5.0,>=2.0.0 in /opt/conda/lib/python3.7/site-p
ackages (from google-auth<2.0dev,>=1.9.0->google-cloud-bigquery==1.25.0) (4.2.2)
Requirement already satisfied: rsa<5,>=3.1.4 in /opt/conda/lib/python3.7/site-packages
(from google-auth<2.0dev,>=1.9.0->google-cloud-bigquery==1.25.0) (4.7.2)
Requirement already satisfied: pyparsing>=2.0.2 in /opt/conda/lib/python3.7/site-package
s (from packaging>=14.3->google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==1.25.0)
(2.4.7)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in /opt/conda/lib/python3.7/site-pac
kages (from pyasn1-modules>=0.2.1->google-auth<2.0dev,>=1.9.0->google-cloud-bigquery==1.
25.0) (0.4.8)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-packages (f
rom requests<3.0.0dev,>=2.18.0->google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==
```

```
1.25.0) (2.10)
Requirement already satisfied: chardet<5,>=3.0.2 in /opt/conda/lib/python3.7/site-packag
es (from requests<3.0.0dev,>=2.18.0->google-api-core<2.0dev,>=1.15.0->google-cloud-bigqu
ery==1.25.0) (4.0.0)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.7/site-packa
ges (from requests<3.0.0dev,>=2.18.0->google-api-core<2.0dev,>=1.15.0->google-cloud-bigq
uery==1.25.0) (2020.12.5)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/conda/lib/python3.7/site-pa
ckages (from requests<3.0.0dev,>=2.18.0->google-api-core<2.0dev,>=1.15.0->google-cloud-b
igquery==1.25.0) (1.26.4)
Installing collected packages: google-resumable-media, google-cloud-bigquery
ERROR: pip's dependency resolver does not currently take into account all the packages t
hat are installed. This behaviour is the source of the following dependency conflicts.
google-cloud-storage 1.38.0 requires google-resumable-media<2.0dev,>=1.2.0, but you have
google-resumable-media 0.5.1 which is incompatible.
Successfully installed google-cloud-bigquery-1.25.0 google-resumable-media-0.5.1
```

Kindly ignore the deprecation warnings and incompatibility errors related to google-cloud-storage.

In [3]:

```
!pip install --user apache-beam[gcp]==2.16.0
!pip install --user httplib2==0.12.0
```

```
Collecting apache-beam[gcp]==2.16.0
  Downloading apache_beam-2.16.0-cp37-cp37m-manylinux1_x86_64.whl (3.0 MB)
     |████████████████████████████████| 3.0 MB 20.2 MB/s eta 0:00:01
Requirement already satisfied: mock<3.0.0,>=1.0.1 in /opt/conda/lib/python3.7/site-packa
ges (from apache-beam[gcp]==2.16.0) (2.0.0)
Requirement already satisfied: pymongo<4.0.0,>=3.8.0 in /opt/conda/lib/python3.7/site-pa
ckages (from apache-beam[gcp]==2.16.0) (3.11.4)
Collecting dill<0.3.1,>=0.3.0
  Downloading dill-0.3.0.tar.gz (151 kB)
     |████████████████████████████████| 151 kB 61.2 MB/s eta 0:00:01
Collecting fastavro<0.22,>=0.21.4
  Downloading fastavro-0.21.24-cp37-cp37m-manylinux1_x86_64.whl (1.2 MB)
     |████████████████████████████████| 1.2 MB 71.1 MB/s eta 0:00:01
Requirement already satisfied: grpcio<2,>=1.12.1 in /opt/conda/lib/python3.7/site-packag
es (from apache-beam[gcp]==2.16.0) (1.37.1)
Requirement already satisfied: crcmod<2.0,>=1.7 in /opt/conda/lib/python3.7/site-package
s (from apache-beam[gcp]==2.16.0) (1.7)
Collecting oauth2client<4,>=2.0.1
  Downloading oauth2client-3.0.0.tar.gz (77 kB)
     |████████████████████████████████| 77 kB 7.7 MB/s  eta 0:00:01
Collecting pyarrow<0.15.0,>=0.11.1
  Downloading pyarrow-0.14.1-cp37-cp37m-manylinux2010_x86_64.whl (58.1 MB)
     |████████████████████████████████| 58.1 MB 5.9 kB/s  eta 0:00:01
| 14.8 MB 61.0 MB/s eta 0:00:01
Requirement already satisfied: hdfs<3.0.0,>=2.1.0 in /opt/conda/lib/python3.7/site-packa
ges (from apache-beam[gcp]==2.16.0) (2.6.0)
Requirement already satisfied: python-dateutil<3,>=2.8.0 in /opt/conda/lib/python3.7/sit
e-packages (from apache-beam[gcp]==2.16.0) (2.8.1)
Collecting httplib2<=0.12.0,>=0.8
  Downloading httplib2-0.12.0.tar.gz (218 kB)
     |████████████████████████████████| 218 kB 71.1 MB/s eta 0:00:01
Requirement already satisfied: pytz>=2018.3 in /opt/conda/lib/python3.7/site-packages (f
rom apache-beam[gcp]==2.16.0) (2021.1)
Collecting pyyaml<4.0.0,>=3.12
  Downloading PyYAML-3.13.tar.gz (270 kB)
     |████████████████████████████████| 270 kB 62.2 MB/s eta 0:00:01
Requirement already satisfied: avro-python3<2.0.0,>=1.8.1 in /opt/conda/lib/python3.7/si
te-packages (from apache-beam[gcp]==2.16.0) (1.9.2.1)
Requirement already satisfied: future<1.0.0,>=0.16.0 in /opt/conda/lib/python3.7/site-pa
ckages (from apache-beam[gcp]==2.16.0) (0.18.2)
Requirement already satisfied: pydot<2,>=1.2.0 in /opt/conda/lib/python3.7/site-packages
(from apache-beam[gcp]==2.16.0) (1.4.2)
```

```
Requirement already satisfied: protobuf<4,>=3.5.0.post1 in /opt/conda/lib/python3.7/site
-packages (from apache-beam[gcp]==2.16.0) (3.16.0)
Collecting google-cloud-pubsub<1.1.0,>=0.39.0
  Downloading google_cloud_pubsub-1.0.2-py2.py3-none-any.whl (118 kB)
     |████████████████████████████████| 118 kB 48.2 MB/s eta 0:00:01
Requirement already satisfied: google-cloud-core<2,>=0.28.1 in /opt/conda/lib/python3.7/
site-packages (from apache-beam[gcp]==2.16.0) (1.6.0)
Collecting google-cloud-datastore<1.8.0,>=1.7.1
  Downloading google_cloud_datastore-1.7.4-py2.py3-none-any.whl (82 kB)
     |████████████████████████████████| 82 kB 1.4 MB/s  eta 0:00:01
Collecting google-apitools<0.5.29,>=0.5.28
  Downloading google-apitools-0.5.28.tar.gz (172 kB)
     |████████████████████████████████| 172 kB 66.7 MB/s eta 0:00:01
Collecting google-cloud-bigtable<1.1.0,>=0.31.1
  Downloading google_cloud_bigtable-1.0.0-py2.py3-none-any.whl (232 kB)
     |████████████████████████████████| 232 kB 54.8 MB/s eta 0:00:01
Collecting cachetools<4,>=3.1.0
  Downloading cachetools-3.1.1-py2.py3-none-any.whl (11 kB)
Collecting google-cloud-bigquery<1.18.0,>=1.6.0
  Downloading google_cloud_bigquery-1.17.1-py2.py3-none-any.whl (142 kB)
     |████████████████████████████████| 142 kB 82.8 MB/s eta 0:00:01
Requirement already satisfied: fasteners>=0.14 in /opt/conda/lib/python3.7/site-packages
(from google-apitools<0.5.29,>=0.5.28->apache-beam[gcp]==2.16.0) (0.16)
Requirement already satisfied: six>=1.12.0 in /opt/conda/lib/python3.7/site-packages (fr
om google-apitools<0.5.29,>=0.5.28->apache-beam[gcp]==2.16.0) (1.16.0)
Collecting google-resumable-media<0.5.0dev,>=0.3.1
  Downloading google_resumable_media-0.4.1-py2.py3-none-any.whl (38 kB)
Requirement already satisfied: google-api-core[grpc]<2.0.0dev,>=1.14.0 in /opt/conda/li
b/python3.7/site-packages (from google-cloud-bigtable<1.1.0,>=0.31.1->apache-beam[gcp]==
2.16.0) (1.26.3)
Requirement already satisfied: grpc-google-iam-v1<0.13dev,>=0.12.3 in /opt/conda/lib/pyt
hon3.7/site-packages (from google-cloud-bigtable<1.1.0,>=0.31.1->apache-beam[gcp]==2.16.
0) (0.12.3)
Requirement already satisfied: google-auth<2.0dev,>=1.21.1 in /opt/conda/lib/python3.7/s
ite-packages (from google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cloud-bigtable<1.1.0,
>=0.31.1->apache-beam[gcp]==2.16.0) (1.30.0)
Requirement already satisfied: setuptools>=40.3.0 in /opt/conda/lib/python3.7/site-packa
ges (from google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cloud-bigtable<1.1.0,>=0.31.1-
>apache-beam[gcp]==2.16.0) (49.6.0.post20210108)
Requirement already satisfied: googleapis-common-protos<2.0dev,>=1.6.0 in /opt/conda/li
b/python3.7/site-packages (from google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cloud-bi
gtable<1.1.0,>=0.31.1->apache-beam[gcp]==2.16.0) (1.53.0)
Requirement already satisfied: requests<3.0.0dev,>=2.18.0 in /opt/conda/lib/python3.7/si
te-packages (from google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cloud-bigtable<1.1.0,>
=0.31.1->apache-beam[gcp]==2.16.0) (2.25.1)
Requirement already satisfied: packaging>=14.3 in /opt/conda/lib/python3.7/site-packages
(from google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cloud-bigtable<1.1.0,>=0.31.1->apa
che-beam[gcp]==2.16.0) (20.9)
Requirement already satisfied: rsa<5,>=3.1.4 in /opt/conda/lib/python3.7/site-packages
(from google-auth<2.0dev,>=1.21.1->google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cloud
-bigtable<1.1.0,>=0.31.1->apache-beam[gcp]==2.16.0) (4.7.2)
Requirement already satisfied: pyasn1-modules>=0.2.1 in /opt/conda/lib/python3.7/site-pa
ckages (from google-auth<2.0dev,>=1.21.1->google-api-core[grpc]<2.0.0dev,>=1.14.0->googl
e-cloud-bigtable<1.1.0,>=0.31.1->apache-beam[gcp]==2.16.0) (0.2.7)
Requirement already satisfied: docopt in /opt/conda/lib/python3.7/site-packages (from hd
fs<3.0.0,>=2.1.0->apache-beam[gcp]==2.16.0) (0.6.2)
Requirement already satisfied: pbr>=0.11 in /opt/conda/lib/python3.7/site-packages (from
mock<3.0.0,>=1.0.1->apache-beam[gcp]==2.16.0) (5.6.0)
Requirement already satisfied: pyasn1>=0.1.7 in /opt/conda/lib/python3.7/site-packages
(from oauth2client<4,>=2.0.1->apache-beam[gcp]==2.16.0) (0.4.8)
Requirement already satisfied: pyparsing>=2.0.2 in /opt/conda/lib/python3.7/site-package
s (from packaging>=14.3->google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cloud-bigtable<
1.1.0,>=0.31.1->apache-beam[gcp]==2.16.0) (2.4.7)
Requirement already satisfied: numpy>=1.14 in /opt/conda/lib/python3.7/site-packages (fr
om pyarrow<0.15.0,>=0.11.1->apache-beam[gcp]==2.16.0) (1.19.5)
```

```
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.7/site-packa
ges (from requests<3.0.0dev,>=2.18.0->google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cl
oud-bigtable<1.1.0,>=0.31.1->apache-beam[gcp]==2.16.0) (2020.12.5)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-packages (f
rom requests<3.0.0dev,>=2.18.0->google-api-core[grpc]<2.0.0dev,>=1.14.0->google-cloud-bi
gtable<1.1.0,>=0.31.1->apache-beam[gcp]==2.16.0) (2.10)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/conda/lib/python3.7/site-pa
ckages (from requests<3.0.0dev,>=2.18.0->google-api-core[grpc]<2.0.0dev,>=1.14.0->google
-cloud-bigtable<1.1.0,>=0.31.1->apache-beam[gcp]==2.16.0) (1.26.4)
Requirement already satisfied: chardet<5,>=3.0.2 in /opt/conda/lib/python3.7/site-packag
es (from requests<3.0.0dev,>=2.18.0->google-api-core[grpc]<2.0.0dev,>=1.14.0->google-clo
ud-bigtable<1.1.0,>=0.31.1->apache-beam[gcp]==2.16.0) (4.0.0)
Building wheels for collected packages: dill, google-apitools, httplib2, oauth2client, p
yyaml
  Building wheel for dill (setup.py) ... done
  Created wheel for dill: filename=dill-0.3.0-py3-none-any.whl size=77512 sha256=8f3d441
59ddc5c37bf848fc193047128fa413d4aaf7a1a62abebe39c17c4baed
  Stored in directory: /home/jupyter/.cache/pip/wheels/6a/3c/26/1fcc712c80b81fe1859f2dda
4415f180fe9ef3ebe9f5e202e4
  Building wheel for google-apitools (setup.py) ... done
  Created wheel for google-apitools: filename=google_apitools-0.5.28-py3-none-any.whl si
ze=130110 sha256=a1ec30e6beecfbe339ac8ba930ebecebefe5ccc84d3023dfe61a5344faa58d53
  Stored in directory: /home/jupyter/.cache/pip/wheels/34/3b/69/ecd8e6ae89d9d71102a58962
c29faa7a9467ba45f99f205920
  Building wheel for httplib2 (setup.py) ... done
  Created wheel for httplib2: filename=httplib2-0.12.0-py3-none-any.whl size=93465 sha25
6=fc92ab0d7a2dfbac10c3d888d9fa7e6aa34c8ebb7932f0cf93047f802bcca5c0
  Stored in directory: /home/jupyter/.cache/pip/wheels/0d/e7/b6/0dd30343ceca921cfbd91f35
5041bd9c69e0f40b49f25b7b8a
  Building wheel for oauth2client (setup.py) ... done
  Created wheel for oauth2client: filename=oauth2client-3.0.0-py3-none-any.whl size=1063
81 sha256=c07a355ac5928e25c96584fbdaf91c9a091e3dd9ee6c92a77d0f464aa7b2ce6c
  Stored in directory: /home/jupyter/.cache/pip/wheels/86/73/7a/3b3f76a2142176605ff38fbc
a574327962c71e25a43197a4c1
  Building wheel for pyyaml (setup.py) ... done
  Created wheel for pyyaml: filename=PyYAML-3.13-cp37-cp37m-linux_x86_64.whl size=43088
sha256=51403a7cb7e675b13f86688bed7ff015c496f7f7c515457b4fe65f68912545f5
  Stored in directory: /home/jupyter/.cache/pip/wheels/95/cd/14/899edaa9cdb9a65aa7224539
f6e0ad488e9a7b202bb48f6ae6
Successfully built dill google-apitools httplib2 oauth2client pyyaml
Installing collected packages: cachetools, httplib2, pyyaml, pyarrow, oauth2client, goog
le-resumable-media, fastavro, dill, google-cloud-pubsub, google-cloud-datastore, google-
cloud-bigtable, google-cloud-bigquery, google-apitools, apache-beam
  WARNING: The script plasma_store is installed in '/home/jupyter/.local/bin' which is n
ot on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use
--no-warn-script-location.
  Attempting uninstall: google-resumable-media
    Found existing installation: google-resumable-media 0.5.1
    Uninstalling google-resumable-media-0.5.1:
      Successfully uninstalled google-resumable-media-0.5.1
  WARNING: The script fastavro is installed in '/home/jupyter/.local/bin' which is not o
n PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use
--no-warn-script-location.
  Attempting uninstall: google-cloud-bigquery
    Found existing installation: google-cloud-bigquery 1.25.0
    Uninstalling google-cloud-bigquery-1.25.0:
      Successfully uninstalled google-cloud-bigquery-1.25.0
  WARNING: The script gen_client is installed in '/home/jupyter/.local/bin' which is not
on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use
--no-warn-script-location.
ERROR: pip's dependency resolver does not currently take into account all the packages t
hat are installed. This behaviour is the source of the following dependency conflicts.
```

```
witwidget 1.7.0 requires oauth2client>=4.1.3, but you have oauth2client 3.0.0 which is i
ncompatible.
tfx 0.28.0 requires apache-beam[gcp]<3,>=2.28, but you have apache-beam 2.16.0 which is
 incompatible.
tfx 0.28.0 requires attrs<21,>=19.3.0, but you have attrs 21.2.0 which is incompatible.
tfx 0.28.0 requires docker<5,>=4.1, but you have docker 5.0.0 which is incompatible.
tfx 0.28.0 requires google-api-python-client<2,>=1.7.8, but you have google-api-python-c
lient 2.3.0 which is incompatible.
tfx 0.28.0 requires kubernetes<12,>=10.0.1, but you have kubernetes 12.0.1 which is inco
mpatible.
tfx 0.28.0 requires pyarrow<3,>=1, but you have pyarrow 0.14.1 which is incompatible.
tfx-bsl 0.28.1 requires apache-beam[gcp]<3,>=2.28, but you have apache-beam 2.16.0 which
is incompatible.
tfx-bsl 0.28.1 requires google-api-python-client<2,>=1.7.11, but you have google-api-pyt
hon-client 2.3.0 which is incompatible.
tfx-bsl 0.28.1 requires pyarrow<3,>=1, but you have pyarrow 0.14.1 which is incompatibl
e.
tensorflow-transform 0.28.0 requires apache-beam[gcp]<3,>=2.28, but you have apache-beam
2.16.0 which is incompatible.
tensorflow-transform 0.28.0 requires pyarrow<3,>=1, but you have pyarrow 0.14.1 which is
incompatible.
tensorflow-model-analysis 0.28.0 requires apache-beam[gcp]<3,>=2.28, but you have apache
-beam 2.16.0 which is incompatible.
tensorflow-model-analysis 0.28.0 requires pyarrow<3,>=1, but you have pyarrow 0.14.1 whi
ch is incompatible.
tensorflow-data-validation 0.28.0 requires apache-beam[gcp]<3,>=2.28, but you have apach
e-beam 2.16.0 which is incompatible.
tensorflow-data-validation 0.28.0 requires joblib<0.15,>=0.12, but you have joblib 1.0.1
which is incompatible.
tensorflow-data-validation 0.28.0 requires pyarrow<3,>=1, but you have pyarrow 0.14.1 wh
ich is incompatible.
pandas-profiling 3.0.0 requires PyYAML>=5.0.0, but you have pyyaml 3.13 which is incompa
tible.
libcst 0.3.18 requires pyyaml>=5.2, but you have pyyaml 3.13 which is incompatible.
google-cloud-storage 1.38.0 requires google-resumable-media<2.0dev,>=1.2.0, but you have
google-resumable-media 0.4.1 which is incompatible.
google-auth-httplib2 0.1.0 requires httplib2>=0.15.0, but you have httplib2 0.12.0 which
is incompatible.
google-api-python-client 2.3.0 requires httplib2<1dev,>=0.15.0, but you have httplib2 0.
12.0 which is incompatible.
cloud-tpu-client 0.10 requires google-api-python-client==1.8.0, but you have google-api-
python-client 2.3.0 which is incompatible.
Successfully installed apache-beam-2.16.0 cachetools-3.1.1 dill-0.3.0 fastavro-0.21.24 g
oogle-apitools-0.5.28 google-cloud-bigquery-1.17.1 google-cloud-bigtable-1.0.0 google-cl
oud-datastore-1.7.4 google-cloud-pubsub-1.0.2 google-resumable-media-0.4.1 httplib2-0.1
2.0 oauth2client-3.0.0 pyarrow-0.14.1 pyyaml-3.13
Requirement already satisfied: httplib2==0.12.0 in /home/jupyter/.local/lib/python3.7/si
te-packages (0.12.0)
```

NOTE: In the output of the above cell you may ignore any WARNINGS or ERRORS related to the following: "apache-beam", "pyarrow", "tensorflow-transform", "tensorflow-model-analysis", "tensorflow-data-validation", "joblib", "google-cloud-storage" etc.

If you get any related errors mentioned above please rerun the above cell.

Note: Restart your kernel to use updated packages.

In [4]:

```
import tensorflow as tf
import apache_beam as beam
import shutil
print(tf.__version__)
```

```
2.4.1
```

1. Environment variables for project and bucket

1. Your project id is the *unique* string that identifies your project (not the project name). You can find this from the GCP Console dashboard's Home page. My dashboard reads: **Project ID:** cloud-training-demos
2. Cloud training often involves saving and restoring model files. Therefore, we should **create a single-region bucket**. If you don't have a bucket already, I suggest that you create one from the GCP console (because it will dynamically check whether the bucket name you want is available) **Change the cell below** to reflect your Project ID and bucket name.

In [8]:
```
import os
PROJECT = 'qwiklabs-gcp-01-1154942ad6ab'    # CHANGE THIS
BUCKET = 'qwiklabs-gcp-01-1154942ad6ab' # REPLACE WITH YOUR BUCKET NAME. Use a regional
REGION = 'us-central1' # Choose an available region for Cloud AI Platform
```

In [9]:
```
# for bash
os.environ['PROJECT'] = PROJECT
os.environ['BUCKET'] = BUCKET
os.environ['REGION'] = REGION
os.environ['TFVERSION'] = '2.1'

## ensure we're using python3 env
os.environ['CLOUDSDK_PYTHON'] = 'python3'
```

In [10]:
```
%%bash
gcloud config set project $PROJECT
gcloud config set compute/region $REGION

## ensure we predict locally with our current Python environment
gcloud config set ml_engine/local_python `which python`
```

```
Updated property [core/project].
Updated property [compute/region].
Updated property [ml_engine/local_python].
```

2. Specifying query to pull the data

Let's pull out a few extra columns from the timestamp.

In [ ]:
```
def create_query(phase, EVERY_N):
  if EVERY_N == None:
    EVERY_N = 4 #use full dataset

  #select and pre-process fields
  base_query = """
SELECT
  (tolls_amount + fare_amount) AS fare_amount,
  DAYOFWEEK(pickup_datetime) AS dayofweek,
  HOUR(pickup_datetime) AS hourofday,
  pickup_longitude AS pickuplon,
  pickup_latitude AS pickuplat,
  dropoff_longitude AS dropofflon,
```

```
  dropoff_latitude AS dropofflat,
  passenger_count*1.0 AS passengers,
  CONCAT(STRING(pickup_datetime), STRING(pickup_longitude), STRING(pickup_latitude), ST
FROM
  [nyc-tlc:yellow.trips]
WHERE
  trip_distance > 0
  AND fare_amount >= 2.5
  AND pickup_longitude > -78
  AND pickup_longitude < -70
  AND dropoff_longitude > -78
  AND dropoff_longitude < -70
  AND pickup_latitude > 37
  AND pickup_latitude < 45
  AND dropoff_latitude > 37
  AND dropoff_latitude < 45
  AND passenger_count > 0
  """

  #add subsampling criteria by modding with hashkey
  if phase == 'train':
    query = "{} AND ABS(HASH(pickup_datetime)) % {} < 2".format(base_query,EVERY_N)
  elif phase == 'valid':
    query = "{} AND ABS(HASH(pickup_datetime)) % {} == 2".format(base_query,EVERY_N)
  elif phase == 'test':
    query = "{} AND ABS(HASH(pickup_datetime)) % {} == 3".format(base_query,EVERY_N)
  return query

print(create_query('valid', 100)) #example query using 1% of data
```

Try the query above in https://bigquery.cloud.google.com/table/nyc-tlc:yellow.trips if you want to see what it does (ADD LIMIT 10 to the query!)

3. Preprocessing Dataflow job from BigQuery

This code reads from BigQuery and saves the data as-is on Google Cloud Storage. We can do additional preprocessing and cleanup inside Dataflow, but then we'll have to remember to repeat that prepreprocessing during inference. It is better to use tf.transform which will do this book-keeping for you, or to do preprocessing within your TensorFlow model. We will look at this in future notebooks. For now, we are simply moving data from BigQuery to CSV using Dataflow.

While we could read from BQ directly from TensorFlow (See: https://www.tensorflow.org/api_docs/python/tf/contrib/cloud/BigQueryReader), it is quite convenient to export to CSV and do the training off CSV. Let's use Dataflow to do this at scale.

Because we are running this on the Cloud, you should go to the GCP Console (https://console.cloud.google.com/dataflow) to look at the status of the job. It will take several minutes for the preprocessing job to launch.

In [13]:

```
%%bash
if gsutil ls | grep -q gs://${BUCKET}/taxifare/ch4/taxi_preproc/; then
  gsutil -m rm -rf gs://$BUCKET/taxifare/ch4/taxi_preproc/
fi
```

First, let's define a function for preprocessing the data

In [14]:

```
import datetime

####
# Arguments:
#   -rowdict: Dictionary. The beam bigquery reader returns a PCollection in
#     which each row is represented as a python dictionary
# Returns:
#   -rowstring: a comma separated string representation of the record with dayofweek
#     converted from int to string (e.g. 3 --> Tue)
####
def to_csv(rowdict):
  days = ['null', 'Sun', 'Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat']
  CSV_COLUMNS = 'fare_amount,dayofweek,hourofday,pickuplon,pickuplat,dropofflon,dropoff
  rowdict['dayofweek'] = days[rowdict['dayofweek']]
  rowstring = ','.join([str(rowdict[k]) for k in CSV_COLUMNS])
  return rowstring


####
# Arguments:
#   -EVERY_N: Integer. Sample one out of every N rows from the full dataset.
#     Larger values will yield smaller sample
#   -RUNNER: 'DirectRunner' or 'DataflowRunner'. Specfy to run the pipeline
#     locally or on Google Cloud respectively.
# Side-effects:
#   -Creates and executes dataflow pipeline.
#     See https://beam.apache.org/documentation/programming-guide/#creating-a-pipeline
####
def preprocess(EVERY_N, RUNNER):
  job_name = 'preprocess-taxifeatures' + '-' + datetime.datetime.now().strftime('%y%m%d
  print('Launching Dataflow job {} ... hang on'.format(job_name))
  OUTPUT_DIR = 'gs://{0}/taxifare/ch4/taxi_preproc/'.format(BUCKET)

  #dictionary of pipeline options
  options = {
    'staging_location': os.path.join(OUTPUT_DIR, 'tmp', 'staging'),
    'temp_location': os.path.join(OUTPUT_DIR, 'tmp'),
    'job_name': 'preprocess-taxifeatures' + '-' + datetime.datetime.now().strftime('%y%
    'project': PROJECT,
    'runner': RUNNER,
    'num_workers' : 4,
    'max_num_workers' : 5
  }
  #instantiate PipelineOptions object using options dictionary
  opts = beam.pipeline.PipelineOptions(flags=[], **options)
  #instantantiate Pipeline object using PipelineOptions
  with beam.Pipeline(options=opts) as p:
      for phase in ['train', 'valid']:
        query = create_query(phase, EVERY_N)
        outfile = os.path.join(OUTPUT_DIR, '{}.csv'.format(phase))
        (
          p | 'read_{}'.format(phase) >> beam.io.Read(beam.io.BigQuerySource(query=quer
            | 'tocsv_{}'.format(phase) >> beam.Map(to_csv)
            | 'write_{}'.format(phase) >> beam.io.Write(beam.io.WriteToText(outfile))
        )
  print("Done")
```

Now, let's run pipeline locally. This takes upto **5 minutes**. You will see a message "Done" when it is done.

In [ ]:
```
preprocess(50*10000, 'DirectRunner')
```

In [ ]:
```
%%bash
gsutil ls gs://$BUCKET/taxifare/ch4/taxi_preproc/
```

4. Run Beam pipeline on Cloud Dataflow

Run pipeline on cloud on a larger sample size.

In [ ]:
```
%%bash
if gsutil ls | grep -q gs://${BUCKET}/taxifare/ch4/taxi_preproc/; then
  gsutil -m rm -rf gs://$BUCKET/taxifare/ch4/taxi_preproc/
fi
```

The following step will take **15-20 minutes.** Monitor job progress on the Cloud Console in the Dataflow section

In [ ]:
```
preprocess(50*100, 'DataflowRunner')
```

Once the job completes, observe the files created in Google Cloud Storage

In [ ]:
```
%%bash
gsutil ls -l gs://$BUCKET/taxifare/ch4/taxi_preproc/
```

In [ ]:
```
%%bash
#print first 10 lines of first shard of train.csv
gsutil cat "gs://$BUCKET/taxifare/ch4/taxi_preproc/train.csv-00000-of-*" | head
```

5. Develop model with new inputs

Download the first shard of the preprocessed data to enable local development.

In [ ]:
```
%%bash
if [ -d sample ]; then
  rm -rf sample
fi
mkdir sample
gsutil cat "gs://$BUCKET/taxifare/ch4/taxi_preproc/train.csv-00000-of-*" > sample/train
gsutil cat "gs://$BUCKET/taxifare/ch4/taxi_preproc/valid.csv-00000-of-*" > sample/valid
```

We have two new inputs in the INPUT_COLUMNS, three engineered features, and the estimator involves bucketization and feature crosses.

In [ ]:
```
%%bash
```

```
grep -A 20 "INPUT_COLUMNS =" taxifare/trainer/model.py
```

In [ ]:
```
%%bash
grep -A 50 "build_estimator" taxifare/trainer/model.py
```

In [ ]:
```
%%bash
grep -A 15 "add_engineered(" taxifare/trainer/model.py
```

Try out the new model on the local sample (this takes **5 minutes**) to make sure it works fine.

In [ ]:
```
%%bash
rm -rf taxifare.tar.gz taxi_trained
export PYTHONPATH=${PYTHONPATH}:${PWD}/taxifare
python -m trainer.task \
  --train_data_paths=${PWD}/sample/train.csv \
  --eval_data_paths=${PWD}/sample/valid.csv  \
  --output_dir=${PWD}/taxi_trained \
  --train_steps=10 \
  --job-dir=/tmp
```

In [ ]:
```
%%bash
ls taxi_trained/export/exporter/
```

You can use `saved_model_cli` to look at the exported signature. Note that the model doesn't need any of the engineered features as inputs. It will compute latdiff, londiff, euclidean from the provided inputs, thanks to the `add_engineered` call in the serving_input_fn.

In [ ]:
```
%%bash
model_dir=$(ls ${PWD}/taxi_trained/export/exporter | tail -1)
saved_model_cli show --dir ${PWD}/taxi_trained/export/exporter/${model_dir} --all
```

In [ ]:
```
%%writefile /tmp/test.json
{"dayofweek": "Sun", "hourofday": 17, "pickuplon": -73.885262, "pickuplat": 40.773008,
```

In [ ]:
```
%%bash
model_dir=$(ls ${PWD}/taxi_trained/export/exporter)
gcloud ai-platform local predict \
  --model-dir=${PWD}/taxi_trained/export/exporter/${model_dir} \
  --json-instances=/tmp/test.json
```

6. Train on cloud

This will take **10-15 minutes** even though the prompt immediately returns after the job is submitted. Monitor job progress on the Cloud Console, in the AI Platform section and wait for the training job to complete.

In [ ]:
```
%%bash
```

```
OUTDIR=gs://${BUCKET}/taxifare/ch4/taxi_trained
JOBNAME=lab4a_$(date -u +%y%m%d_%H%M%S)
echo $OUTDIR $REGION $JOBNAME
gsutil -m rm -rf $OUTDIR
gcloud ai-platform jobs submit training $JOBNAME \
  --region=$REGION \
  --module-name=trainer.task \
  --package-path=${PWD}/taxifare/trainer \
  --job-dir=$OUTDIR \
  --staging-bucket=gs://$BUCKET \
  --scale-tier=BASIC \
  --runtime-version 2.1 \
  --python-version 3.5 \
  -- \
  --train_data_paths="gs://$BUCKET/taxifare/ch4/taxi_preproc/train*" \
  --eval_data_paths="gs://${BUCKET}/taxifare/ch4/taxi_preproc/valid*"  \
  --train_steps=5000 \
  --output_dir=$OUTDIR
```

The RMSE is now 8.33249, an improvement over the 9.3 that we were getting ... of course, we won't know until we train/validate on a larger dataset. Still, this is promising. But before we do that, let's do hyper-parameter tuning.

Use the Cloud Console link to monitor the job and wait till the job is done.