feature_engineering

November 25, 2021

1 Feature Engineering

---

1. Import packages
2. Load data
3. Feature engineering

1.1 1. Import packages

[ ]:
```
import pandas as pd
```

1.2 2. Load data

[ ]:
```
df = pd.read_csv('./clean_data_after_eda.csv')
df["date_activ"] = pd.to_datetime(df["date_activ"], format='%Y-%m-%d')
df["date_end"] = pd.to_datetime(df["date_end"], format='%Y-%m-%d')
df["date_modif_prod"] = pd.to_datetime(df["date_modif_prod"], format='%Y-%m-%d')
df["date_renewal"] = pd.to_datetime(df["date_renewal"], format='%Y-%m-%d')
```

[ ]:
```
df.head(3)
```

[ ]:
```
                                  id                    channel_sales  \
0  24011ae4ebbe3035111d65fa7c15bc57  foosdfpfkusacimwkcsosbicdxkicaua
1  d29c2c54acc38ff3c0614d0a653813dd                           MISSING
2  764c75f661154dac3a6c254cd082ea7d  foosdfpfkusacimwkcsosbicdxkicaua

   cons_12m  cons_gas_12m  cons_last_month date_activ   date_end  \
0         0         54946                0 2013-06-15 2016-06-15
1      4660             0                0 2009-08-21 2016-08-30
2       544             0                0 2010-04-16 2016-04-16

  date_modif_prod date_renewal  forecast_cons_12m  …  \
0      2015-11-01   2015-06-23               0.00  …
1      2009-08-21   2015-08-31             189.95  …
```

```
2      2010-04-16   2015-04-17              47.96  …

   var_6m_price_off_peak_var  var_6m_price_peak_var  \
0                   0.000131           4.100838e-05
1                   0.000003           1.217891e-03
2                   0.000004           9.450150e-08

   var_6m_price_mid_peak_var  var_6m_price_off_peak_fix  \
0                   0.000908                   2.086294
1                   0.000000                   0.009482
2                   0.000000                   0.000000

   var_6m_price_peak_fix  var_6m_price_mid_peak_fix var_6m_price_off_peak  \
0              99.530517                  44.235794              2.086425
1               0.000000                   0.000000              0.009485
2               0.000000                   0.000000              0.000004

   var_6m_price_peak  var_6m_price_mid_peak  churn
0       9.953056e+01               44.236702      1
1       1.217891e-03                0.000000      0
2       9.450150e-08                0.000000      0

[3 rows x 44 columns]
```

1.3 3. Feature engineering

1.3.1 Difference between off-peak prices in December and preceding January

Below is the code created by your colleague to calculate the feature described above. Use this code to re-create this feature and then think about ways to build on this feature to create features with a higher predictive power.

[ ]:
```
price_df = pd.read_csv('price_data.csv')
price_df["price_date"] = pd.to_datetime(price_df["price_date"],␣
 ↪format='%Y-%m-%d')
price_df.head()
```

[ ]:
```
                                 id price_date  price_off_peak_var  \
0  038af19179925da21a25619c5a24b745 2015-01-01            0.151367
1  038af19179925da21a25619c5a24b745 2015-02-01            0.151367
2  038af19179925da21a25619c5a24b745 2015-03-01            0.151367
3  038af19179925da21a25619c5a24b745 2015-04-01            0.149626
4  038af19179925da21a25619c5a24b745 2015-05-01            0.149626

   price_peak_var  price_mid_peak_var  price_off_peak_fix  price_peak_fix  \
0             0.0                 0.0           44.266931             0.0
```

```
1            0.0                0.0           44.266931            0.0
2            0.0                0.0           44.266931            0.0
3            0.0                0.0           44.266931            0.0
4            0.0                0.0           44.266931            0.0

   price_mid_peak_fix
0                 0.0
1                 0.0
2                 0.0
3                 0.0
4                 0.0
```

[ ]:
```
# Group off-peak prices by companies and month
monthly_price_by_id = price_df.groupby(['id', 'price_date']).
 ↪agg({'price_off_peak_var': 'mean', 'price_off_peak_fix': 'mean'}).
 ↪reset_index()

# Get january and december prices
jan_prices = monthly_price_by_id.groupby('id').first().reset_index()
dec_prices = monthly_price_by_id.groupby('id').last().reset_index()

# Calculate the difference
diff = pd.merge(dec_prices.rename(columns={'price_off_peak_var': 'dec_1',␣
 ↪'price_off_peak_fix': 'dec_2'}), jan_prices.drop(columns='price_date'),␣
 ↪on='id')
diff['offpeak_diff_dec_january_energy'] = diff['dec_1'] -␣
 ↪diff['price_off_peak_var']
diff['offpeak_diff_dec_january_power'] = diff['dec_2'] -␣
 ↪diff['price_off_peak_fix']
diff = diff[['id',␣
 ↪'offpeak_diff_dec_january_energy','offpeak_diff_dec_january_power']]
diff.head()
```

[ ]:
```
                                 id  offpeak_diff_dec_january_energy  \
0  0002203ffbb812588b632b9e628cc38d                        -0.006192
1  0004351ebdd665e6ee664792efc4fd13                        -0.004104
2  0010bcc39e42b3c2131ed2ce55246e3c                         0.050443
3  0010ee3855fdea87602a5b7aba8e42de                        -0.010018
4  00114d74e963e47177db89bc70108537                        -0.003994

   offpeak_diff_dec_january_power
0                        0.162916
1                        0.177779
2                        1.500000
3                        0.162916
4                       -0.000001
```