Untitled10

October 24, 2020

0.0.1 Kaggle Competition

```
[1]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     %matplotlib inline
[2]: data = pd.read_csv("/Users/graceyin/Downloads/phase1_training_data.csv").
      →to_numpy()
[3]: | X_all = data[data[:, 0] == "CA"]
[4]: y_can = X_all[:, 3]
[5]: y_can = X_all[:, 3]
     y_can = np.reshape(y_can, (y_can.shape[0], 1))
     y_can = y_can.astype(float)
[]:
[6]: #retrieve data thats canadian alone
     def get_feature(X, col_name):
         X = X_all[:, col_name]
         X = np.reshape(X, (X.shape[0], 1))
         X = X.astype(float)
         return X
[7]: X_can_all = np.delete(X_all, [0, 1, 3], axis = 1)
     X_can_all = np.reshape(X_can_all, (X_can_all.shape[0], 3))
     X_can_all = X_can_all.astype(float)
[8]: time_series = y_can
[9]: import pandas as pd
     from matplotlib import pylab as plt
     from statsmodels.tsa.arima_model import ARMA
     import seaborn as sns
```

```
import statsmodels.api as sm
import random
import numpy as np
import statsmodels
```

[10]: time_series = y_can

```
[11]: def trend(t, amp=1):
          return amp*(1 + np.sin(t/10))
      #length of time series
      n_{time\_steps} = 250
      #amplitud of time series
      amplitud=10
      noise frac aplitud= 0.5
      #initializing the time series
      time series = np.zeros(n time steps)
      time_series[0] = trend(0, amplitud)
      \#The\ AR(1) parameter. Our goal will be to find this parameter.
      alpha = 0.1
      #making the time series
      for t in range(1,n_time_steps):
          time_series[t] = (1 - alpha)*time_series[t - 1] + alpha*trend(t, ___
       →amp=amplitud) + alpha*np.random.normal(0,noise_frac_aplitud*amplitud)
      #passing the time series to a pandas format
      dates = sm.tsa.datetools.dates_from_range('2000m1', length=len(time_series))
      time_series_pd= pd.Series(time_series, index=dates)
```

```
[12]: window = 40
    n_iter = n_time_steps - window
    alpha_list = []
    alpha_elite_list = []
    #n_time_steps
    for i in range(n_iter):
        #not rolling window but intervals... to fix
        temp_time_series_pd = time_series_pd[i:window + i]
        plt.plot(temp_time_series_pd)
        res = sm.tsa.detrend(temp_time_series_pd, order=2)
        mod = ARMA(res, order=(1,0))
        ar1_fit = mod.fit()

        score = statsmodels.tsa.arima_model.ARMAResults(mod,ar1_fit.params)

#The alpha parameter is...
        alpha_list.append(1 - ar1_fit.params[1])
```

```
if score.pvalues[1] < 0.001:</pre>
        alpha_elite_list.append(1 - ar1_fit.params[1])
print("real alpha = ", alpha)
print("estimated alpha = ", np.mean(alpha_elite_list))
print("standard deviation = ", np.std(alpha_elite_list))
/opt/anaconda3/lib/python3.7/site-
packages/pandas/plotting/_matplotlib/converter.py:103: FutureWarning: Using an
implicitly registered datetime converter for a matplotlib plotting method. The
converter was registered by pandas on import. Future versions of pandas will
require you to explicitly register matplotlib converters.
To register the converters:
        >>> from pandas.plotting import register matplotlib converters
        >>> register_matplotlib_converters()
  warnings.warn(msg, FutureWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-packages/scipy/signal/signaltools.py:1341:
FutureWarning: Using a non-tuple sequence for multidimensional indexing is
deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will
be interpreted as an array index, `arr[np.array(seq)]`, which will result either
in an error or a different result.
  out full[ind] += zi
/opt/anaconda3/lib/python3.7/site-packages/scipy/signal/signaltools.py:1344:
FutureWarning: Using a non-tuple sequence for multidimensional indexing is
deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will
be interpreted as an array index, `arr[np.array(seq)]`, which will result either
in an error or a different result.
  out = out full[ind]
/opt/anaconda3/lib/python3.7/site-packages/scipy/signal/signaltools.py:1350:
FutureWarning: Using a non-tuple sequence for multidimensional indexing is
deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will
be interpreted as an array index, `arr[np.array(seq)]`, which will result either
in an error or a different result.
  zf = out_full[ind]
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
 % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
 % freq, ValueWarning)
```

```
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
 % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/kalmanf/kalmanfilter.py:221: RuntimeWarning: divide by
zero encountered in true_divide
  Z_mat, R_mat, T_mat)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
  % freq, ValueWarning)
/opt/anaconda3/lib/python3.7/site-
packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency
information was provided, so inferred frequency M will be used.
 % freq, ValueWarning)
```

/opt/anaconda3/lib/python3.7/site-packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency information was provided, so inferred frequency M will be used.

% freq, ValueWarning)

/opt/anaconda3/lib/python3.7/site-

packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency information was provided, so inferred frequency M will be used.

% freq, ValueWarning)

/opt/anaconda3/lib/python3.7/site-

packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency information was provided, so inferred frequency M will be used.

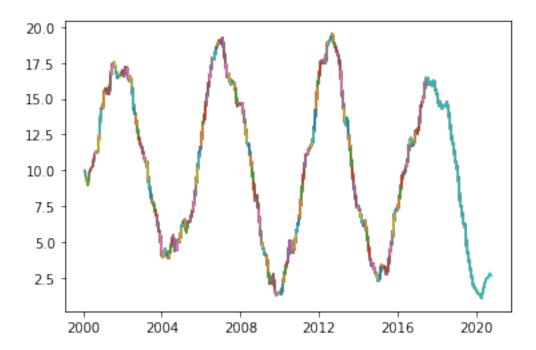
% freq, ValueWarning)

/opt/anaconda3/lib/python3.7/site-

packages/statsmodels/tsa/base/tsa_model.py:165: ValueWarning: No frequency information was provided, so inferred frequency M will be used.

% freq, ValueWarning)

real alpha = 0.1
estimated alpha = 0.1584267886753695
standard deviation = 0.1051980894213107



real alpha = 0.1

```
[14]: def trend(t, amp=1):
          return amp*(1 + np.sin(t/10))
      #length of time series
      n_{time_steps} = 250
      #amplitud of time series
      amplitud=10
      noise_frac_aplitud= 0.5
      #initializing the time series
      time_series = np.zeros(n_time_steps)
      time_series[0] = trend(0, amplitud)
      #The AR(1) parameter. Our goal will be to find this parameter.
      alpha = 0.1
      #making the time series
      for t in range(1,n_time_steps):
          time_series[t] = (1 - alpha)*time_series[t - 1] + alpha*trend(t,__
       →amp=amplitud) + alpha*np.random.normal(0,noise_frac_aplitud*amplitud)
      #passing the time series to a pandas format
      dates = sm.tsa.datetools.dates_from_range('2000m1', length=len(time_series))
      time_series_pd= pd.Series(time_series, index=dates)
[15]: def trend(t, amp=1):
          return amp*(1 + np.sin(t/10))
[16]: trend(0, 250)
[16]: 250.0
 []:
[17]:
              y_can = np.reshape(y_can, (y_can.shape[0], 1))
              y_can = y_can.astype(float)
              #retrieve data thats canadian alone
              def get_feature(X, col_name):
                  X = X_all[:, col_name]
                  X = np.reshape(X, (X.shape[0], 1))
                  X = X.astype(float)
                  return X
              X_{\text{can\_all}} = \text{np.delete}(X_{\text{all}}, [0, 1, 3], \text{ axis } = 1)
              X_can_all = np.reshape(X_can_all, (X_can_all.shape[0], 3))
              X_can_all = X_can_all.astype(float)
```

```
X_can_cases = get_feature(X_all, 2)
              X_can_cases_14_100k = get_feature(X_all, 4)
              X_can_cases_100k = get_feature(X_all, 5)
[13]: y_world = X_world[:, 3]
      y_world = np.reshape(y_world, (y_world.shape[0], 1))
      y_world = y_world.astype(float)
      X_world_cases = X_world[:, 2]
      X_world_cases = np.reshape(X_world_cases, (X_world_cases.shape[0], 1))
      X_world_cases = X_world_cases.astype(float)
             TypeError
                                                        Traceback (most recent call
      →last)
             <ipython-input-13-23cb2675b65a> in <module>
         ----> 1 y_world = X_world[:, 3]
               2 y_world = np.reshape(y_world, (y_world.shape[0], 1))
               3 y_world = y_world.astype(float)
               5 X_world_cases = X_world[:, 2]
             /opt/anaconda3/lib/python3.7/site-packages/pandas/core/frame.py in u
      →__getitem__(self, key)
                             if self.columns.nlevels > 1:
            2978
            2979
                                 return self._getitem_multilevel(key)
         -> 2980
                             indexer = self.columns.get_loc(key)
                             if is_integer(indexer):
            2981
            2982
                                 indexer = [indexer]
             /opt/anaconda3/lib/python3.7/site-packages/pandas/core/indexes/base.pyu
      →in get_loc(self, key, method, tolerance)
            2895
            2896
                             try:
         -> 2897
                                 return self._engine.get_loc(key)
            2898
                             except KeyError:
                                 return self._engine.get_loc(self.
            2899
      →_maybe_cast_indexer(key))
             pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()
```

```
pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

TypeError: '(slice(None, None, None), 3)' is an invalid key

[]:
```