

# Numerai Competition



# Problem statement

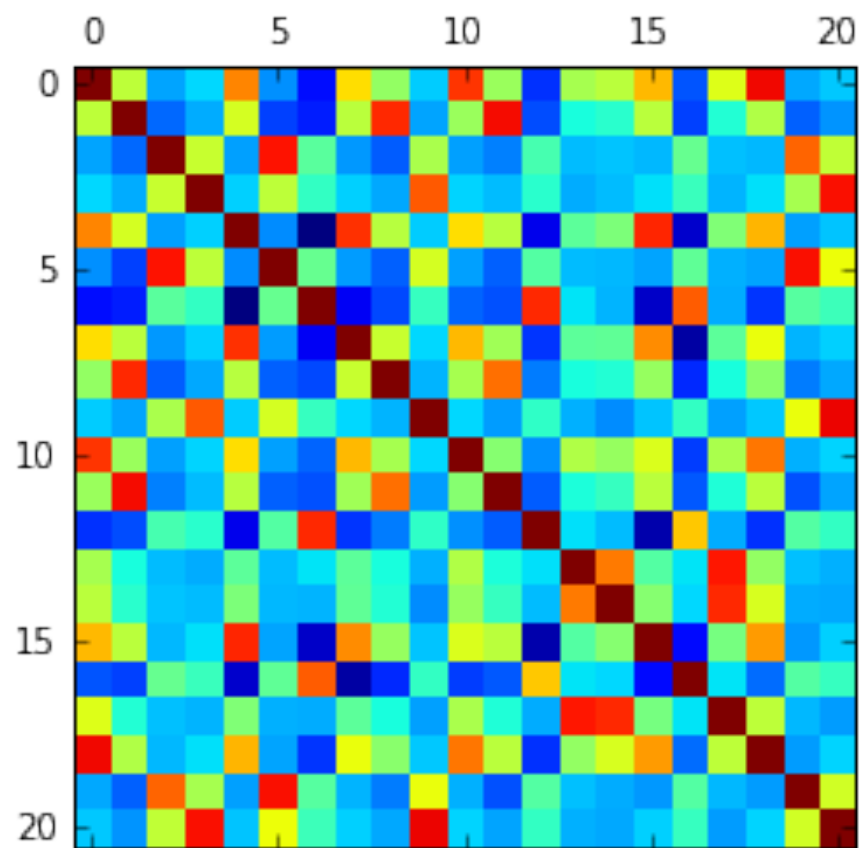
- Perfectly balanced binary classification
- Anonymized 21 features
- Data changes every week

# Data leakage

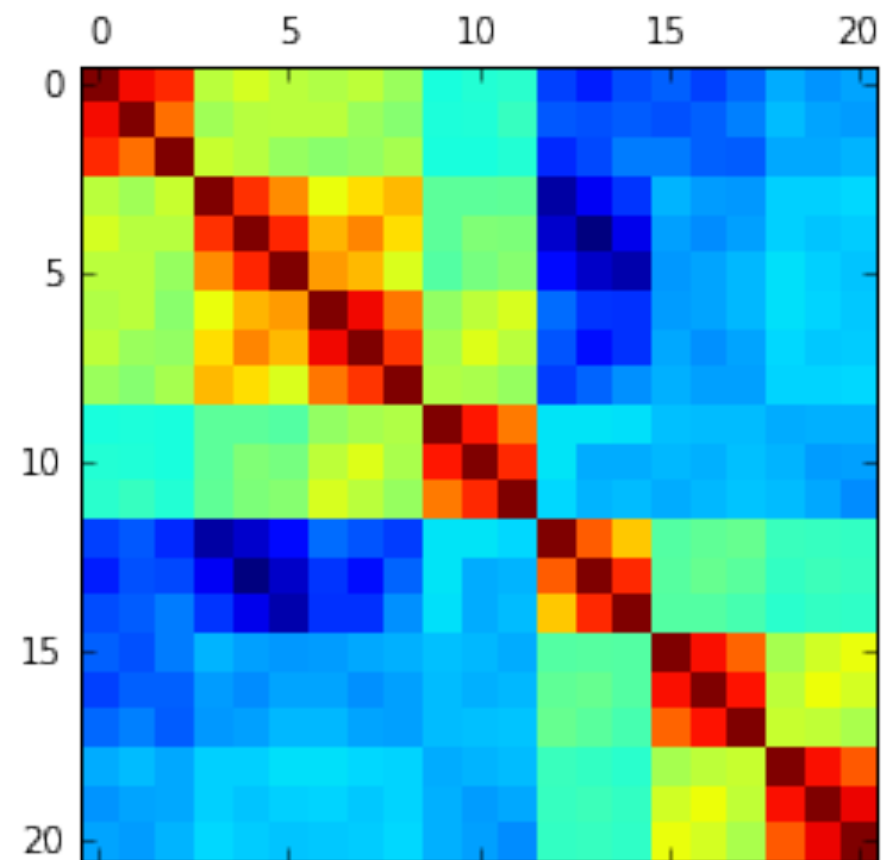
- Allegedly time series
- Target variable depends on changes between each point (think of returns)
- Approximate reconstruction of true order via nearest neighbourhood analysis
- Top 10 via logistic regression on 21 original features + 21 features from nearest neighbour

# Hardcore EDA

Correlation matrix

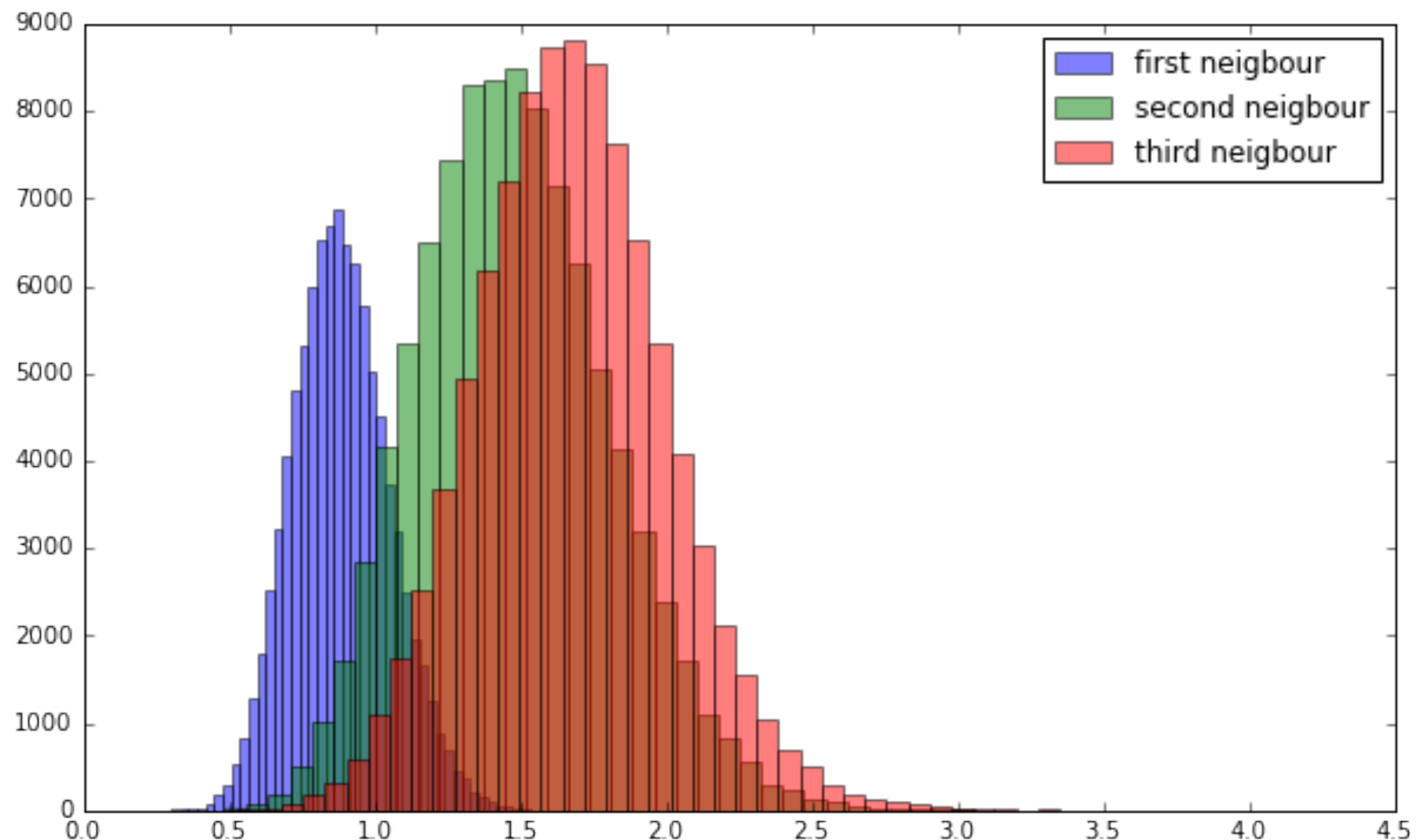


Sorted correlation matrix



# Hardcore EDA

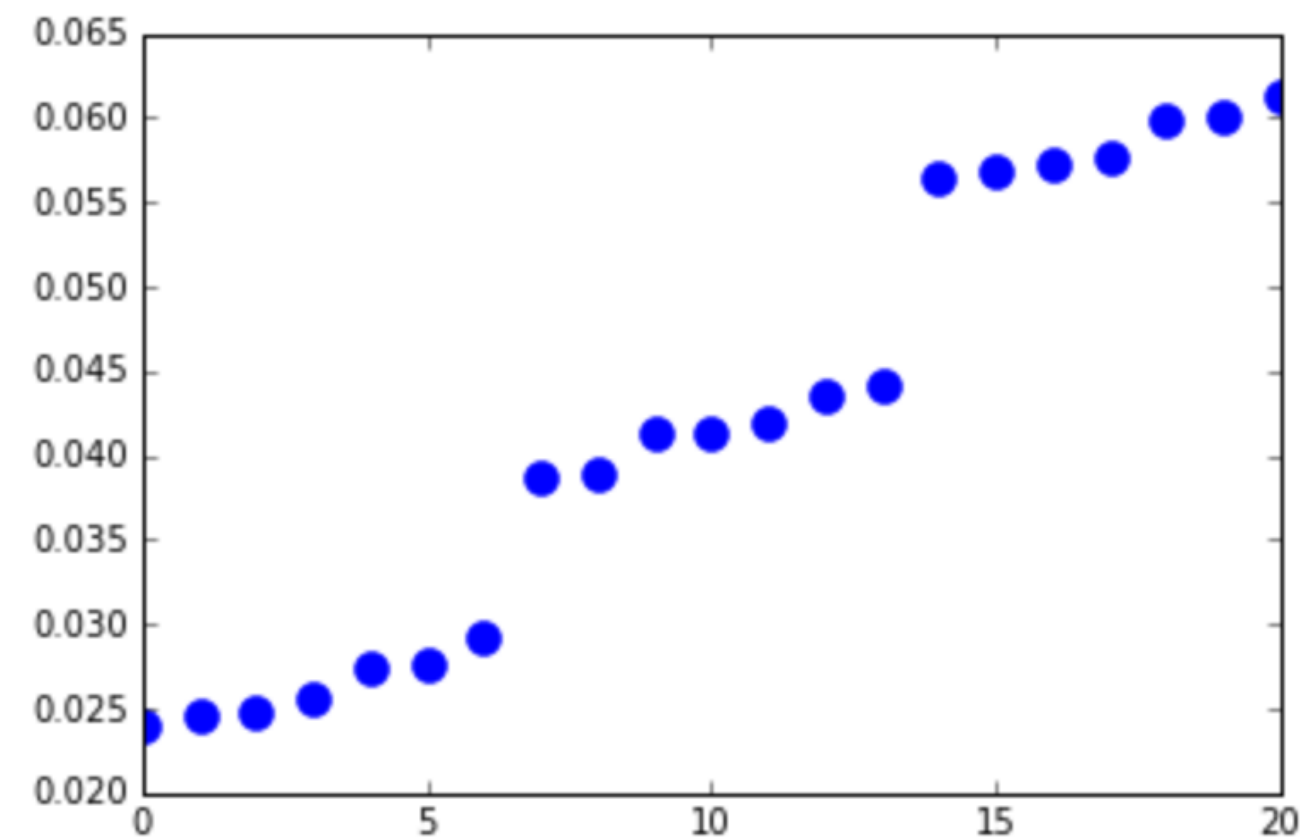
Distribution of distances to the first 3 neighbors in previous weeks data



# Hardcore EDA

```
#inds - indices of nearest neighbours  
ff_groups = [[ 1,  4, 17,  6, 20,  0, 19], [5, 15,  9, 12, 14, 18, 11], [7,  8, 10,  
          13, 16,  3,  2]]  
mm = np.mean(np.abs(train[true_order].values - train_old[true_order].values[inds[:,0]]),axis=0)  
  
plt.plot(sorted(mm), '.', markersize=20)
```

```
[<matplotlib.lines.Line2D at 0x7ff1f189df50>]
```



# New data is a lie



Thank you!