Financial
Data Science

Chartered Financial Data Scientist

Project Work

Steve Huse

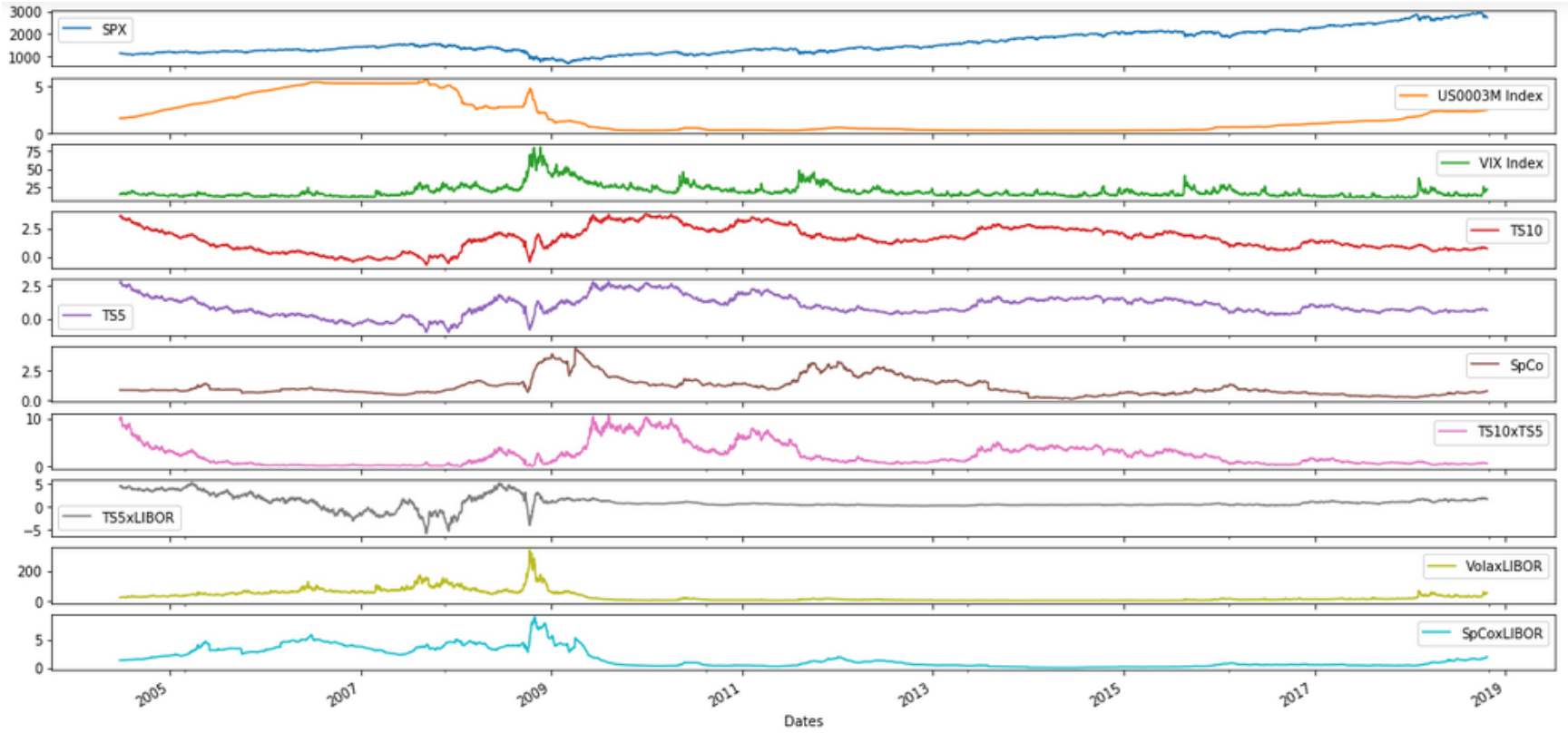
Introduction

- Currently our Portfolio Managers on shares are using various models like CPPI, VolTarget, BMI, TSI, etc. to gather information about markets and manage their investment decisions.
- Objective therefore was to develop an integrated approach estimating market situation, which delivers results as objective as possible.
- Hauptmann et al. (2014): “Forecasting market turbulences using regime-switching models”
- Distinguishing three Market phases (calm, turbulent bullish, turbulent bearish) using Markov-switching models on monthly returns of the S&P 500.
- Replaced these Models and clustered via k-Means on a wider range of market data in order to find some pattern in the data, which reveal different types of market situations.
- Next step is to predict the kind of the forthcoming market phase, which would allow to align the investment strategy.
- For better explanation some linear regression model are calibrated, which should give a better look into the data and their effect on forthcoming period.

Data

- Used Dataset contains time series from 2004 until now on a daily basis for each of the following facts:
  - S&P 500 Index (SPX)
  - 3 Month USD LIBOR (US0003M Index)
  - VIX Index
  - 10Y USD Swap – 3M LIBOR (TS10)
  - 5Y USD Swap – 3M LIBOR (TS5)
  - Spreads between BBB and AAA Corporate Bonds (SpCo)
  - Different Products of the mentioned Factors: (TS10 * TS5), (TS5 * LIBOR), (Vola * LIBOR) and (Vola * SpCo)
- Except the yields from the corporate bonds the data where directly retrieved from Bloomberg.
- As source for the 10Y USD Swap and the 5Y USD Swap USSWAP10 Curncy and USSWAP5 Curncy where used respectively, i.e. the swap rate against the 3M LIBOR.
- The spreads are defined as the difference in between the average yields of the iBoxx Corporate AAA Index and the iBoxx Corporate BBB Index.

Data



- Data was stored and processed in a DataFrame-object from pandas.
- Methods like .fillna and .dropna were used to fill gaps in the data and to drop records which could not be refilled in a proper way.

Data

- Function was defined to compress the data to different periods of time, e.g. weekly, monthly, …
- Adds a column which contains the logreturn of the S&P 500 for the certain period of time.

```
def logreturn (X,iv=1) :

    X=pd.concat([X,pd.DataFrame(np.zeros((len(X),1)),index=X.index,columns=['LogReturn'])], axis=1)
    for i in range(0,len(X)-iv) :
        X.LogReturn[i]=np.log(X['SPX'][i]/X['SPX'][i+iv])
    X_remain=X.iloc[:len(X)-iv:iv,:]
    X_drop=X[~X.index.isin(X_remain.index)]

    return X_remain,X_drop
```

- Following examination is done for a period of 5, i.e. weekly data.
- Data was furthermore divided in a training set from 2004 until the end of 2017 and a test set from beginning of 2018 until now
- Using .StandardScaler from Scikit-Learn the training set was standardized and the transformation gathered this way was also applied to the test set.

Clustering

- In Preparation for the clustering the S&P 500 Index as well as the Logreturn were been removed from the training set.
- As algorithm for clustering k-Means was used to derive three clusters from the data analogous to the mentioned article.
- Clustercenters:

	US0003M Index	VIX Index	TS10	TS5	SpCo	TS10xTS5	TS5xLIBOR	VolaxLIBOR	SpCoxLIBOR
0	0.531332	16.333457	1.692331	0.960848	1.039364	1.833462	0.435163	8.004623	0.482310
1	1.118815	22.213955	2.730579	1.871482	1.571293	5.382987	1.848496	22.644347	1.550502
2	4.472415	19.318989	0.425307	0.171314	1.015780	0.348813	0.367395	78.856682	3.974718

- In addition and for a better visualization of the results from the clustering a principal component analysis for three components was done using PCA from Scikit-Learn.
- Explained Variance:

```
pca.explained_variance_ratio_
```

```
array([0.4962562 , 0.23849509, 0.13691566])
```

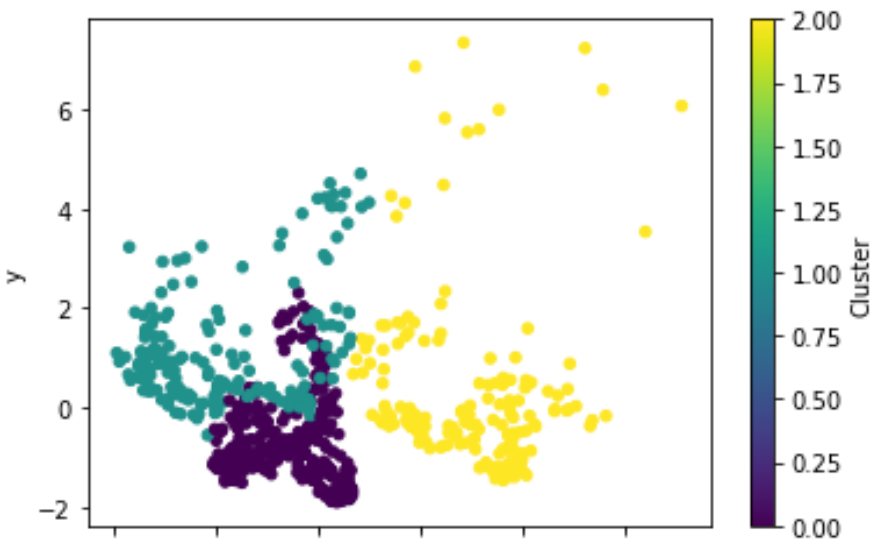
```
sum(pca.explained_variance_ratio_)
```

```
0.8716669482726709
```

Clustering

```
pd.DataFrame(pca.components_.round(4),columns=X_train.columns, index=['x','y','z'])
```

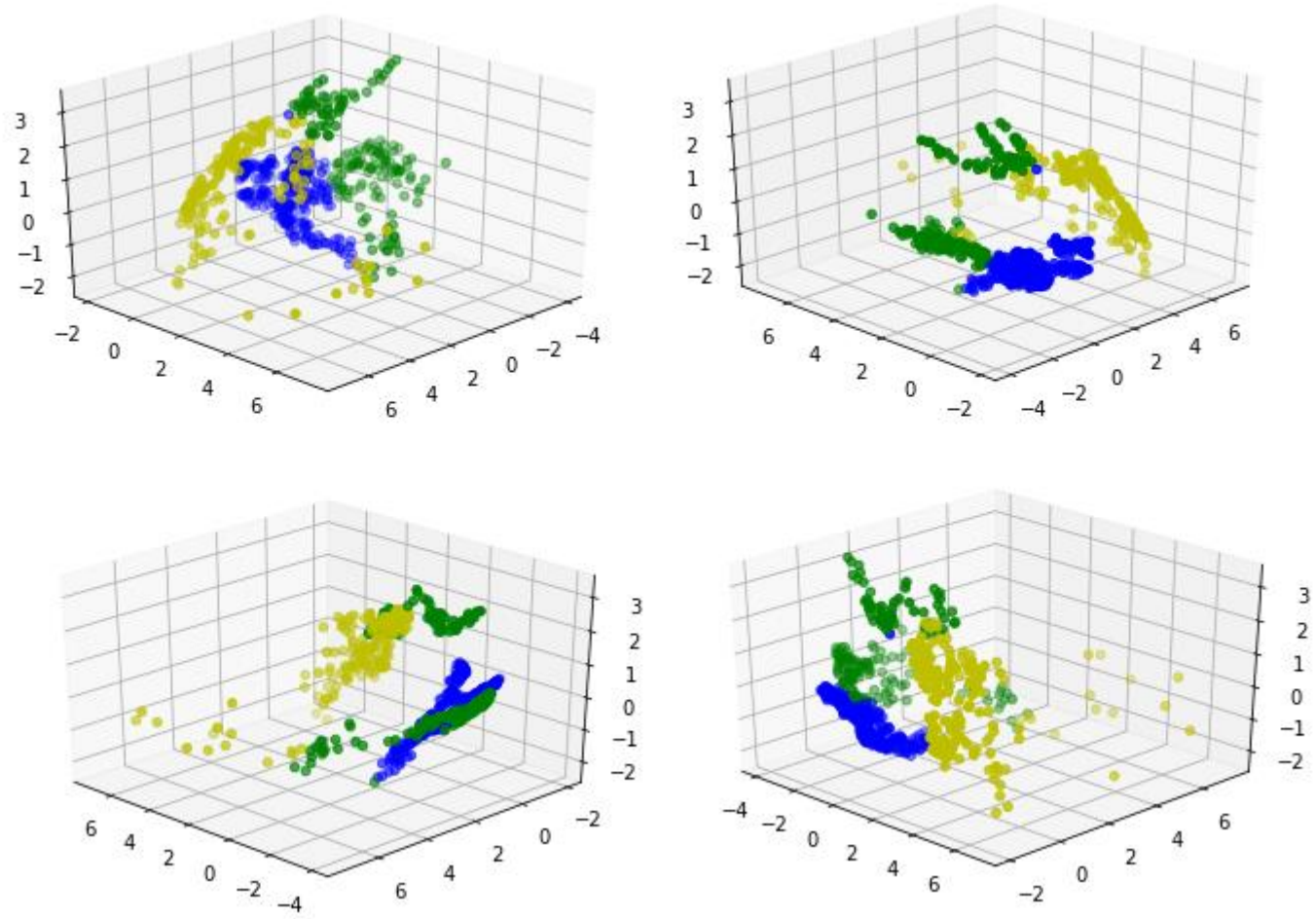
	US0003M Index	VIX Index	TS10	TS5	SpCo	TS10xTS5	TS5xLIBOR	VolaxLIBOR	SpCoxLIBOR
x	0.4063	0.0307	-0.4461	-0.4335	-0.0383	-0.3871	-0.1420	0.3798	0.3623
y	0.0093	0.6060	0.1666	0.1651	0.5439	0.1764	0.1938	0.2880	0.3608
z	0.3666	-0.2866	0.0820	0.2498	-0.3532	0.1871	0.6761	0.1711	0.2612



- The composition of the three components derived from PCA (above)
- Three k-Means clusters shown in the first two components of PCA (left)
- Cluster “2” is separated sharply from the other two clusters in these two components

Clustering

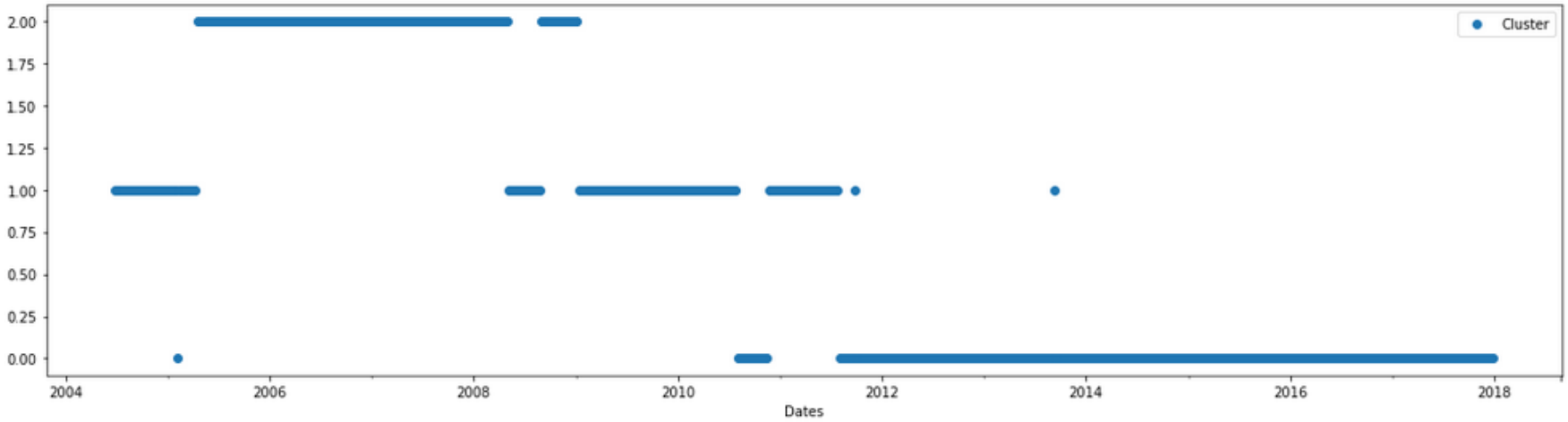
3D-Plot using Matplotlib (Cluster “0” blue, “1” green, “2” yellow):



- Compact Cluster “0”. Deeper Data needed to distinguish in between this Cluster.
- “1” seems to build two groups divided in the “z”-axis. Obviously more information in the data than recognized by k-Means Clusters.

Prediction

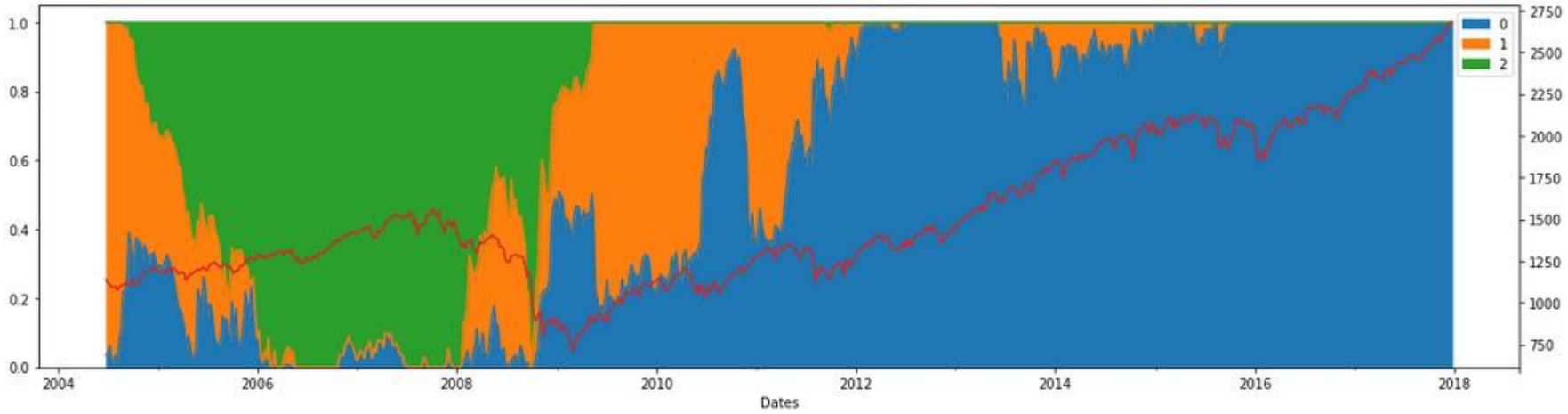
- At this point the clusters look like this in a time-dependent view:



- More meaningful to estimate probabilities to which cluster a data point belongs in a certain point of time.
- Solved by performing k-nearest neighbors algorithm from Scikit-Learn.
- Probabilities of a certain point estimated by looking on the classification of the one fifth nearest of all points.
- Delivers a proper smoothing

Prediction

- Together with the S&P 500 Index it looks like this:

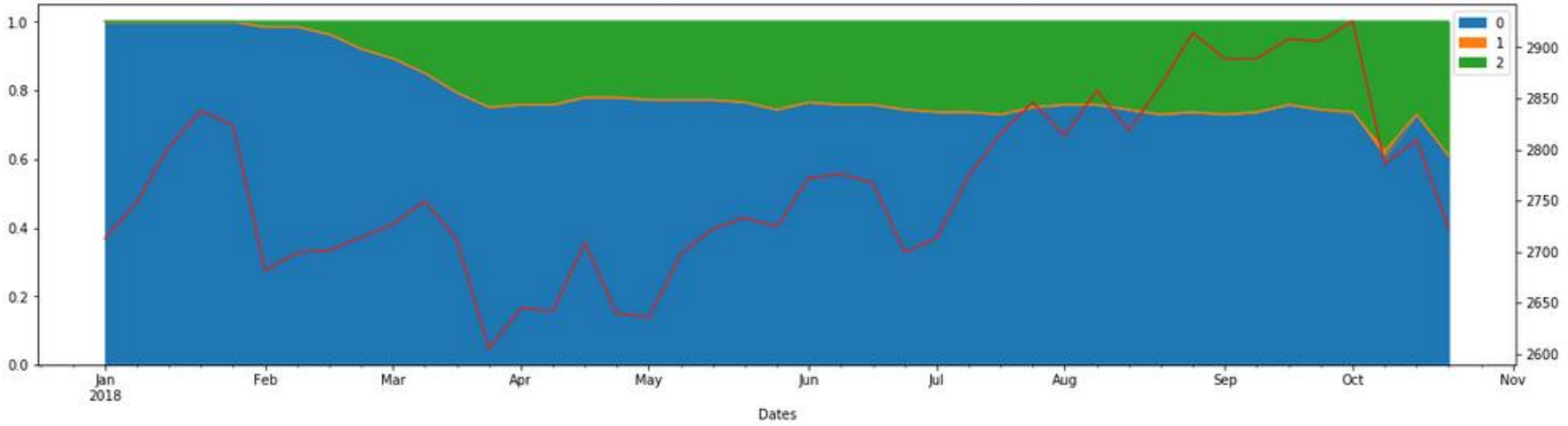


	US0003M Index	VIX Index	TS10	TS5	SpCo	TS10xTS5	TS5xLIBOR	VolaxLIBOR	SpCoxLIBOR
0	0.531332	16.333457	1.692331	0.960848	1.039364	1.833462	0.435163	8.004623	0.482310
1	1.118815	22.213955	2.730579	1.871482	1.571293	5.382987	1.848496	22.644347	1.550502
2	4.472415	19.318989	0.425307	0.171314	1.015780	0.348813	0.367395	78.856682	3.974718

- Keeping in mind that the S&P 500 takes no place in classification there seems to be a surprisingly high correlation to certain movements the index did in the past years

Prediction

- Something even more surprising happens when the estimation is performed on the test data, i.e. the year 2018.



- A probability for cluster “2” appears as it was last seen in times of the financial crisis around the bankruptcy of Lehman Brothers.
- So do we have to bring our money to a safe place?
- Leads to the question about how useful these estimated probabilities are for the prediction of future movements.
- Last step is trying to forecast the probabilities of the next period.

Forecast via linear Regression

- Three linear regression models where calibrated with the training data by shifting the knn-probabilities in the previous period.
- Each model should deliver an estimator for a certain cluster probability in the next period.
- Linear models were chosen because of the good explanatory power of a linear estimator.
- Results:

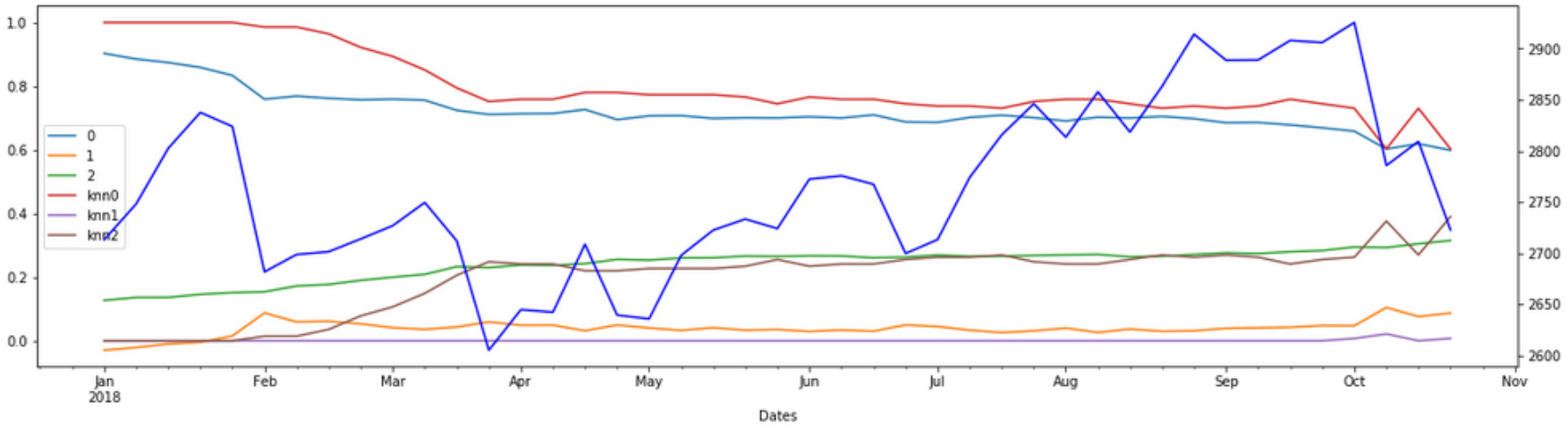
	Intercept	US0003M Index	VIX Index	TS10	TS5	SpCo	TS10xTS5	TS5xLIBOR	VolaxLIBOR	SpCoxLIBOR
0	0.574305	-0.402257	-0.052554	-0.097349	0.189512	-0.017197	-0.306643	-0.073421	0.038910	-0.078711
1	0.183740	0.008456	0.052108	-0.081425	-0.237969	0.033824	0.447711	0.150617	-0.023778	-0.008095
2	0.241955	0.393801	0.000446	0.178774	0.048457	-0.016627	-0.141068	-0.077196	-0.015132	0.086805

	0	1	2
MSE	0.009019	0.001996	0.005002
R^2	0.967602	0.953317	0.973747

- Metrics suggest a very good adaption to the data and a almost linear connection between the input data an the derived cluster probabilities.
- But due to the construction of the probabilities there are no big changes from one period to another.

Forecast via linear Regression

- The application of the estimators on the test data reveals the problem in forecasting this way.
- The estimation seems to forecast future developments at least in the test interval (period from January to April this year).
- But this means not necessarily that the estimators are suitable for prediction of the development of the S&P 500 in a short-term view.



- They deliver maybe another indicator for more general market situations, but at first much more tests have to be done to validate this suggestion.
- In this case they could be useful to explain and manage strategic investment decisions.

Forecast via linear Regression

- Correlation-matrix knn-probabilities to SPX on training data:

	0	1	2	0x1	0x2	1x2	LogReturn	SPX
0	1.000000	-0.435411	-0.796372	-0.123468	-0.429989	-0.464360	0.068483	0.652002
1	-0.435411	1.000000	-0.197718	0.751934	0.094050	0.252840	0.000765	-0.587762
2	-0.796372	-0.197718	1.000000	-0.370732	0.405027	0.335773	-0.075085	-0.315083
0x1	-0.123468	0.751934	-0.370732	1.000000	0.011860	-0.061936	0.019221	-0.491346
0x2	-0.429989	0.094050	0.405027	0.011860	1.000000	0.534927	0.010398	-0.386770
1x2	-0.464360	0.252840	0.335773	-0.061936	0.534927	1.000000	-0.070623	-0.342758
LogReturn	0.068483	0.000765	-0.075085	0.019221	0.010398	-0.070623	1.000000	0.065341
SPX	0.652002	-0.587762	-0.315083	-0.491346	-0.386770	-0.342758	0.065341	1.000000

- Correlation-matrix knn-probabilities to SPX on test data:

	0	1	2	0x1	0x2	1x2	LogReturn	SPX
0	1.000000	-0.386794	-0.999518	-0.385779	-0.982706	-0.387727	0.185955	-0.170793
1	-0.386794	1.000000	0.357974	0.998132	0.238385	0.995724	-0.450717	0.100154
2	-0.999518	0.357974	1.000000	0.357009	0.987004	0.359064	-0.173111	0.169563
0x1	-0.385779	0.998132	0.357009	1.000000	0.239623	0.988263	-0.438502	0.117022
0x2	-0.982706	0.238385	0.987004	0.239623	1.000000	0.236275	-0.106419	0.153341
1x2	-0.387727	0.995724	0.359064	0.988263	0.236275	1.000000	-0.467323	0.073206
LogReturn	0.185955	-0.450717	-0.173111	-0.438502	-0.106419	-0.467323	1.000000	0.285163
SPX	-0.170793	0.100154	0.169563	0.117022	0.153341	0.073206	0.285163	1.000000

Appendix: additional approaches

- Application of logistic regression:
  - Currently the estimator produces values, which can be higher than one and lower than zero. For estimated probabilities this not very suitable.
  - Logistic regression could solve this problem. Not done yet.
- Application of seemingly unrelated regression:
  - So far no taking in to account that the three regression problems are related in sense of that the result must sum up to one.
  - Modified linear regression in this sense was done with little improvement due to already good shape of the models
- Application of simple cluster approach:
  - Looking at the logreturns of the S&P 500 a simple clustering was done, so that ten percent of periods with the highest logreturns build the first cluster, the ten percent with the lowest logreturns the second and every other period was assigned to the third one.
  - The results this way were non-interpretable as they looked quite randomly distributed
- Next to do:
  - Examine the suitability of this approach to other indices, especially some European indices
  - Translation of the used data to equivalent facts in European economics
  - Looking for and examining the impacts of additional facts like the gold price, oil price or other macroeconomic ratios
  - Selection of facts using p-values or other hypothesis tests