**Results**

We have performed analysis to obtain comprehensive result statistics and model performance, as well as drill-in crisis-period-focused examination via the back testing exercise.

we first may focus on the performance of our early warning indicator around the 2008 stock market crash and then around the 1998 one as showcases We then step back to provide the overall statistical summary of the model performances and comparison studies between models with TDA feature bolted on and without.

The period of sharpest decline for the S&P 500 in 2008 was around the period of the collapse of Lehman Brothers. From the middle of September to end of November the S&P declined from 1255 to 800, a decline of 36%. It then recovered somewhat ending 2008 at around 900. Within the September-November period there are two episodes of decline, one until the third week of October where it fell 23%, and then over the month of November where it fell another 10%. As we see from the top panel of Figure 1 our TDA based indicator rises sharply in the middle of September just before the first sharp decline, rising sharply from around 0.3 to above 0.8 in the course of a week. Since the beginning of 2008 its previous peak had been around 0.5 in February. It then drops back sharply for the next three weeks, remaining well above its previous level, and then peaks at around 0.85 in the third week of October and remains above 0.7 for the next two weeks, just before the next decline. It then falls sharply for three weeks, and rises sharply for the first two weeks of December where there is no major decline. From the bottom panel of Figure 1 we see that the non TDA indicator shows an increase in the first week of September to around 0.5, which is at the same levels in February, and then continues to decline until it jumps to just under 0.7 in the first week of November 2008. The overall peak is just under 0.9 in the first week of December, well after the sharpest declines have occurred.

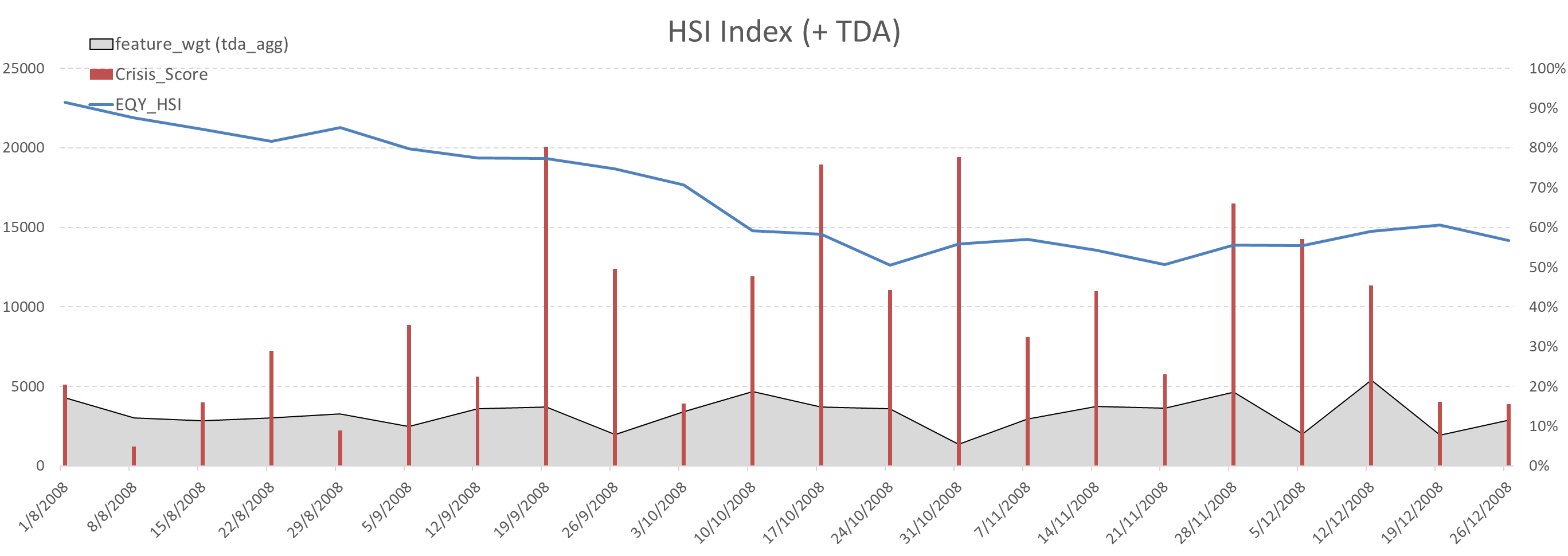
The TDA based indicator’s increase in mid-September was well above anything observed in the course of the year while the non TDA indicator’s sharp jump in early September was still at a level seen earlier in the year when no subsequent decline was observed. It thus seems that the TDA based indicator’s signal would have received more attention. The peak of the non TDA based indicator in December could be regarded as a false positive as could the sharp increase for the TDA based indicator. However the low levels of the non TDA based indicator through the rest of September and all of October, when they were consistently below levels in February, could be regarded as a false negative which is a far more serious issue in the context of financial crash prediction.

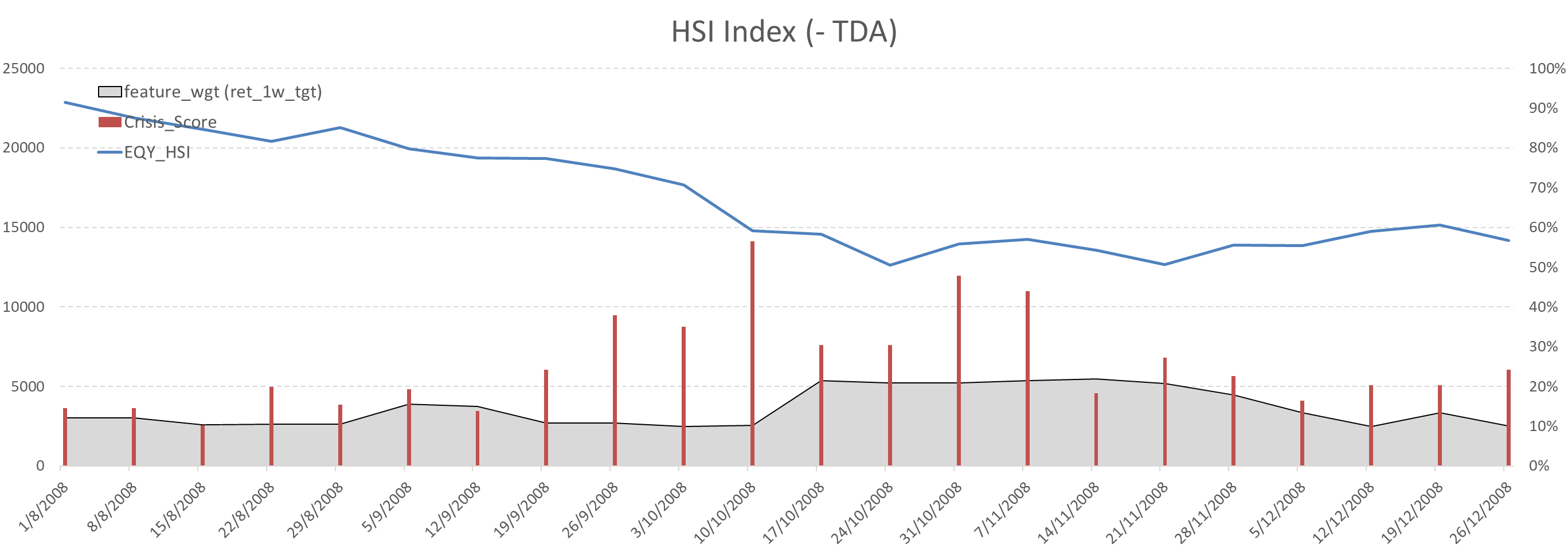
We next analyse the 1998 situation when the S&P declined around 18% from mid-July to early September, another case of a short, sharp decline followed the debt crisis in Russia early of the year. In this case, as we see from the top panel of Figure 2, the TDA based indicator rises sharply to above 0.9 in the first two weeks of July from levels of around 0.3 in the previous four weeks and then continues to decline, remaining at around 0.3 from August until the end of the year. There is a sharp increase to 0.6 at the end of May, which could be interpreted as a short term false positive. The non TDA indicator as we see from the bottom panel of Figure 2, increases sharply at the beginning of June but remains at a very low level throughout July and then peaks at almost 1.0 at the end of August. Both of these peaks, particularly the second one could be regarded as false positives and the low levels of the indicator in July could certainly be regarded as a false negative. Thus for this particular financial crisis, it is clear that the non TDA based indicator is unable to detect the crash while the TDA based indicator does a much better job.

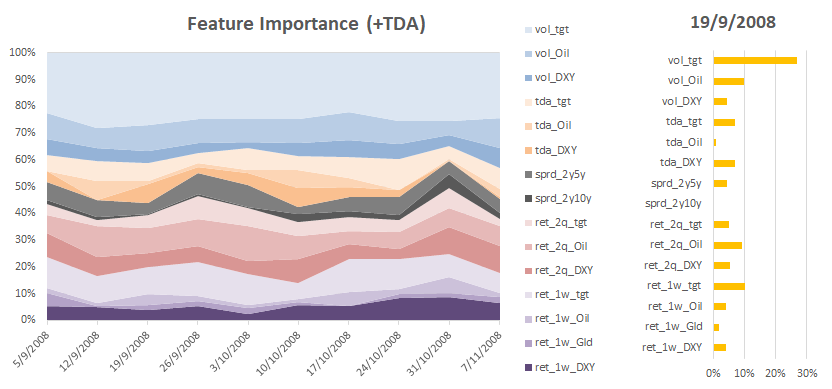
We next analyse the behaviour of both indicators for two other indices, namely the FTSE 100, the major UK index and the Hang Seng, the main Hong Kong index, both over the 2008 crisis period. The sharpest period of decline for the FTSE was in the first week of October and as we see from Figure 2 the TDA based indicator peaks at exactly this point and then peaks again at the end of October which was the beginning of a less sharp decline. The non TDA based indicator’s peak of around 0.7 is during the third and fourth weeks of October near the bottom of the sharp decline. The pattern is similar for the Hang Seng index where the major decline begins around the middle of September. The TDA based indicator jumps around this point, which is in fact the peak of 0.8, and jumps sharply again in mid-October which is the beginning of a shallower decline. There is a third jump at the end of October which is a week before another 10% decline in November, and there are two potential false positives at the end of November. In contrast the non TDA based indicator’s peak is much lower at around 0.6 and is at the end of the first major decline in the second week of October and before the second shallower decline. There is a thus a consistent pattern of both performance of the TDA based indicator in signalling a decline as well in the differences between the performance of the TDA and non TDA based indicators.

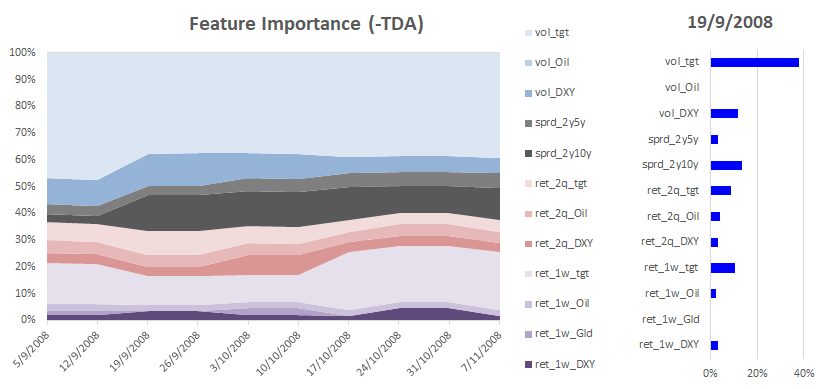
[more paragraphs for HSI and FTSE here preceding the charts?]

Hang Seng Index (with Feature Importance Analysis)

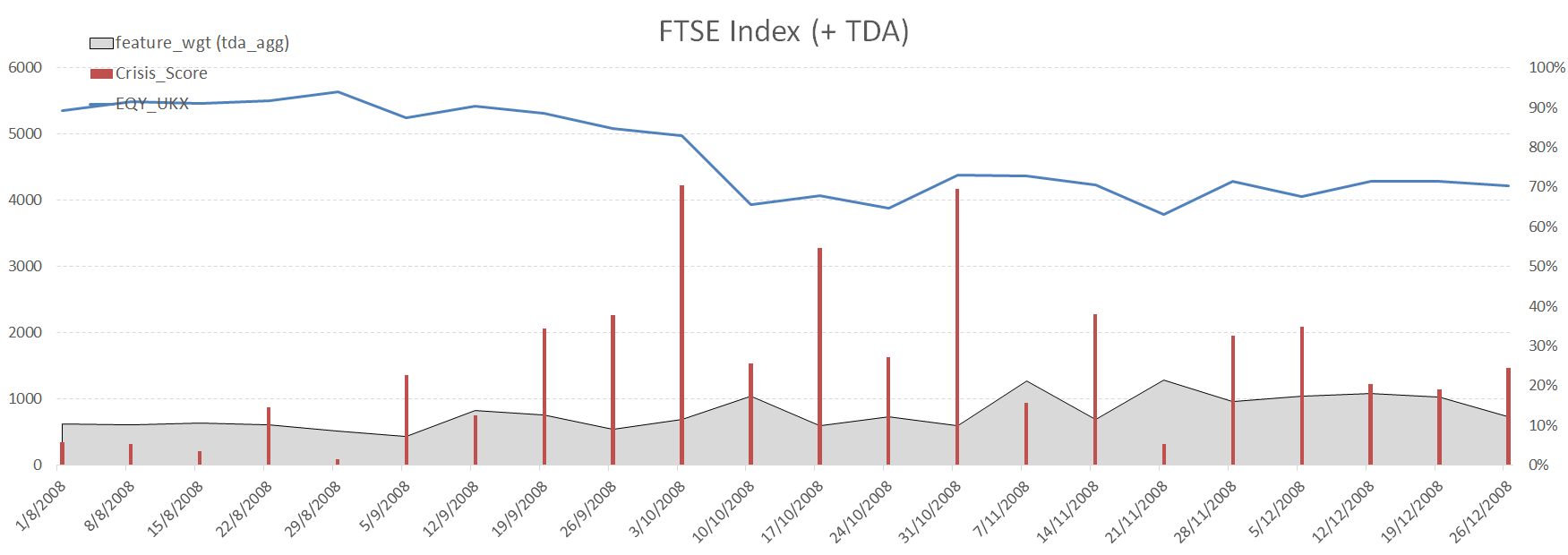


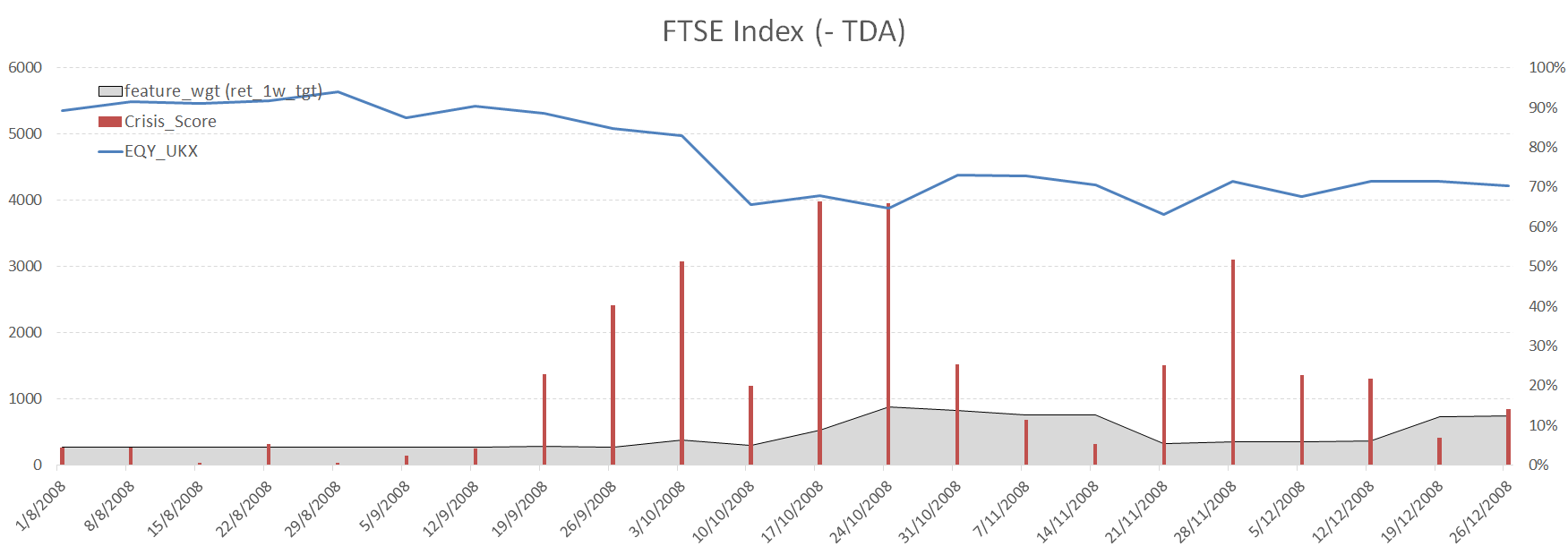


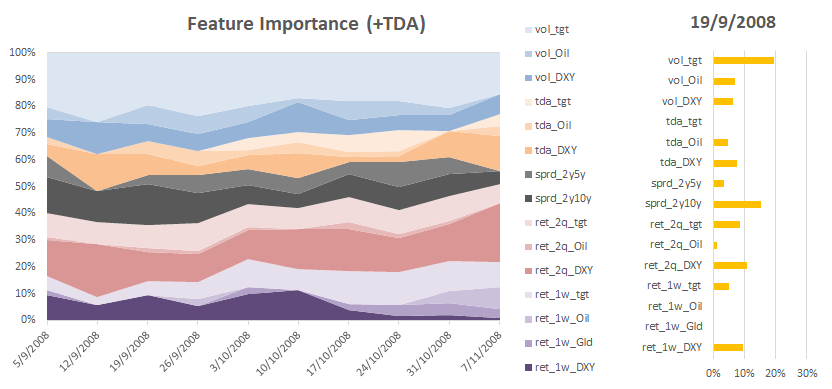


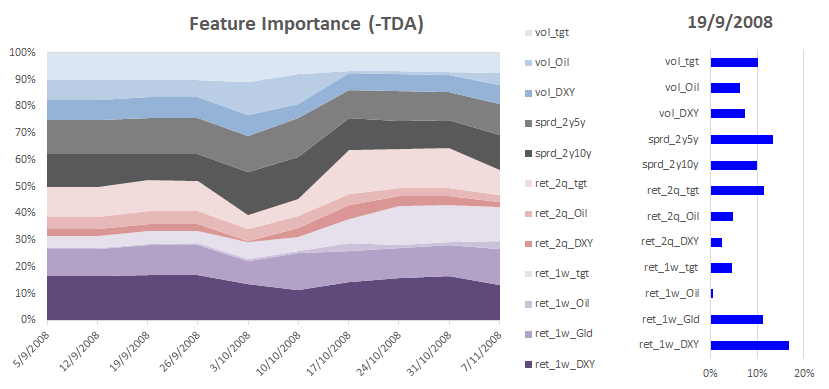


FTSE 100









The nature of the TDA based indicator is quite different from the non TDA based indicator as can be seen from the features and feature weighting. The TDA based indicator has the potential to unlock and better coordinate extra features based on TDA and so has access to a larger set of features. The feature weightings during September to November 2008 for both indicators are shown in Figure 3 and for the non TDA based indicator we see the dominance of two features, S&P volatility and the previous week’s return for gold emerging as the dominant features over this period. The S&P volatility has the highest weighting, between 20% and 30% over this period while that for the lagged gold return is between 15% and 20%. The six month return on the S&P and crude oil also receive weightings of around 10%, with several features receiving no weighting at all. Hence the non TDA indicator seems to weight certain factors quite heavily and some others not at all, suggesting that it tends to optimize on training in sample data quite aggressively which then leaves it susceptible to the problem of overfitting. For both the market declines we focus on, the price of overfitting seems to be both false positives as well as false negatives. In contrast, as we see from Figure 2, the TDA based indicator does not give weightings of over 20% consistently to any of the features, and all features including the TDA features are weighted. In fact only one feature, the S&P volatility consistently receives weightings above 10% over this period. The other features that receive weightings around 10% are the lagged weekly return on the S&P and the six month return on crude oil. The S&P TDA feature receives the highest weighting of the TDA features, around 5%. The TDA based indicator thus does not optimize nearly as aggressively which allows the model training to pick up weaker signals that could contribute to its improved performance relative to the non TDA indicator. Thus incorporating TDA into the machine learning based framework seems to “unlock” a much wider variety of signals and therefore avoid some of the overfitting problems that the non TDA based indicator appears to have. During the 1998 market decline the overall pattern is very similar. For the non TDA based indicator three features, namely crude oil volatility and the lagged weekly return on the S&P as well as gold, all receive weighting above 20% while most of the other features receive zero weighting. In the case of the TDA based indicator only one feature, the volatility of the dollar index, receives weighting of around 20%, while three features of the S&P namely its volatility, lagged one week and six month return all receive weightings of around 10%. All of the other features including the TDA features receive some weighting. Thus while both indicators give high weighting to short and medium term momentum type indicators the TDA indicator also weights other potentially weaker signals which seem to help with crash prediction.

For the FTSE over the 2008 crisis, the non TDA indicator weights a larger number of features than for the S&P with six features receiving weights of around 10% as we see from Figure 4. The TDA based indicator weights four features around 10% with the S&P volatility receiving a weighting close to 20%. The pattern of weightings is somewhat different between the two indicators, with the gold return highly weighted for the non TDA indicator and not at all for the TDA indicator and six month returns more highly weighted for the TDA indicator, for example, suggesting that the TDA indicator tends to focus more on the longer term signals. The pattern for the Hang Seng is more similar to that for the S&P with the non TDA indicator weighting S&P volatility at around 40% and also weighting the lagged one week return on the S&P over 10%. The TDA based indicator also weights S&P volatility highest, but around 25% and weights the other features more uniformly with longer term returns receiving much higher weights than for the non TDA based indicator. At least one TDA feature receives a weighting of around 10% for both the FTSE and the Hang Seng. Thus the pattern of differences in feature weighting with the non TDA indicator weighting short term momentum signals and the TDA indicator weighting longer term, potentially weaker, signals is also visible for these indices.

We next focus on the relative importance of different factors for the S&P 500 over the entire 24 year sample period. We consider three of the most highly weighted features, namely the lagged one week return on the S&P, the volatility of the S&P and the lagged one week return on gold, both with and without incorporating TDA features. In addition we aggregate the weights of the three TDA features and the results are shown in Table 1. A number of stylized features emerge. The weights of the three non-TDA features exhibit considerable time variation across the two halves of the sample period. In the first half the two return based features exhibit much higher average weightings, with the gold return feature having average weightings above 20%, which are higher in the absence of TDA features. However in the second half of the sample the volatility feature dominates and the average weight in the absence of TDA is almost double that in the presence (29% relative to 15%) while the return based features average weighting is below 10%. Thus volatility emerges as the most significant feature in this early warning system since 2006. The same pattern is visible for the TDA based features with their average weighting more than double over the second period compared with the first (17% relative to 8%). Taken together these suggest a possible shift in the nature of the S&P over the second half of our sample.

[more paragraphs for in-depth crisis-period-focused presentation…]

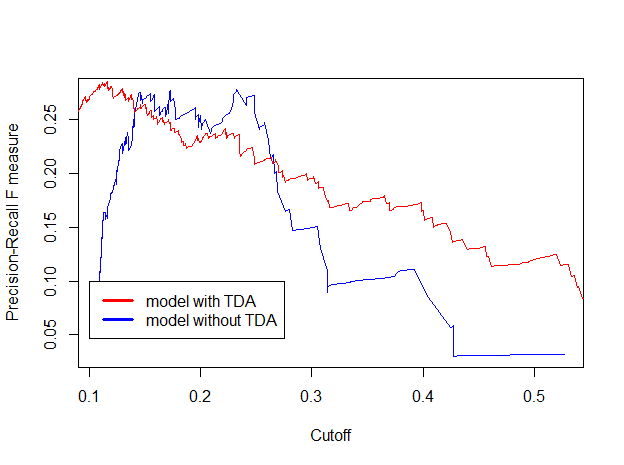
I can then write the statistical results supported by score charts and tables?

The “zoom in” focus on the model performances during crisis periods provides us with intuitive appreciation on how TDA features add value. We now “zoom out” to describe the overall performances and comparison studies for models trained with and without TDA features.

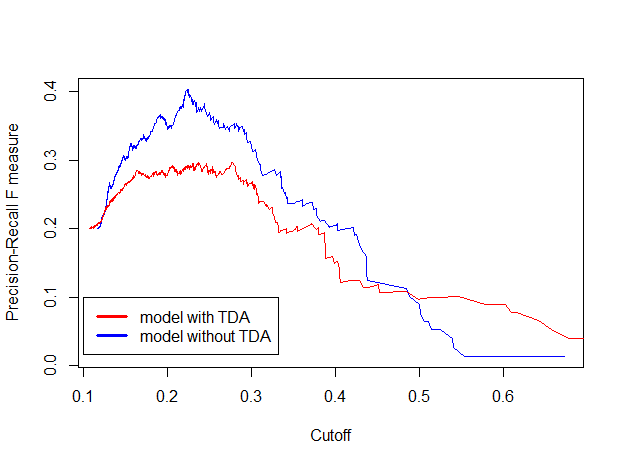
By focusing on the model performances during crisis periods, we can see the boosted accuracy of the TDA models in detecting market crashes when it really is happening, namely, the improvement in reducing false negatives, especially when the bets are high. We aggregate and quantify this observation over the entire backtesting period by the classic F1 score in machine learning practice.

Figure 5 summarises the TDA advantage visually for all three major market indices we have tested.

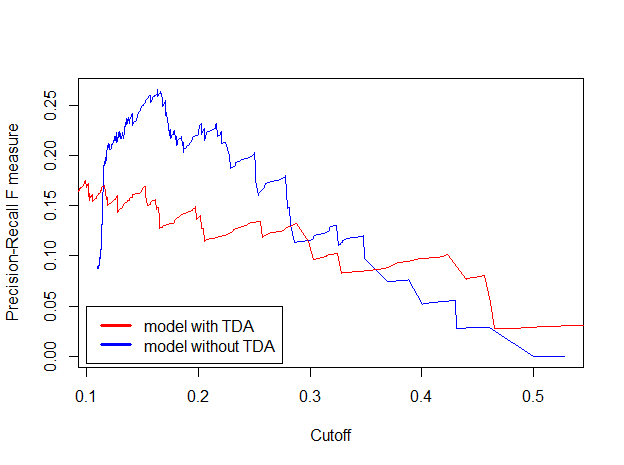
For S&P 500 Index:



For Hang Seng Index:



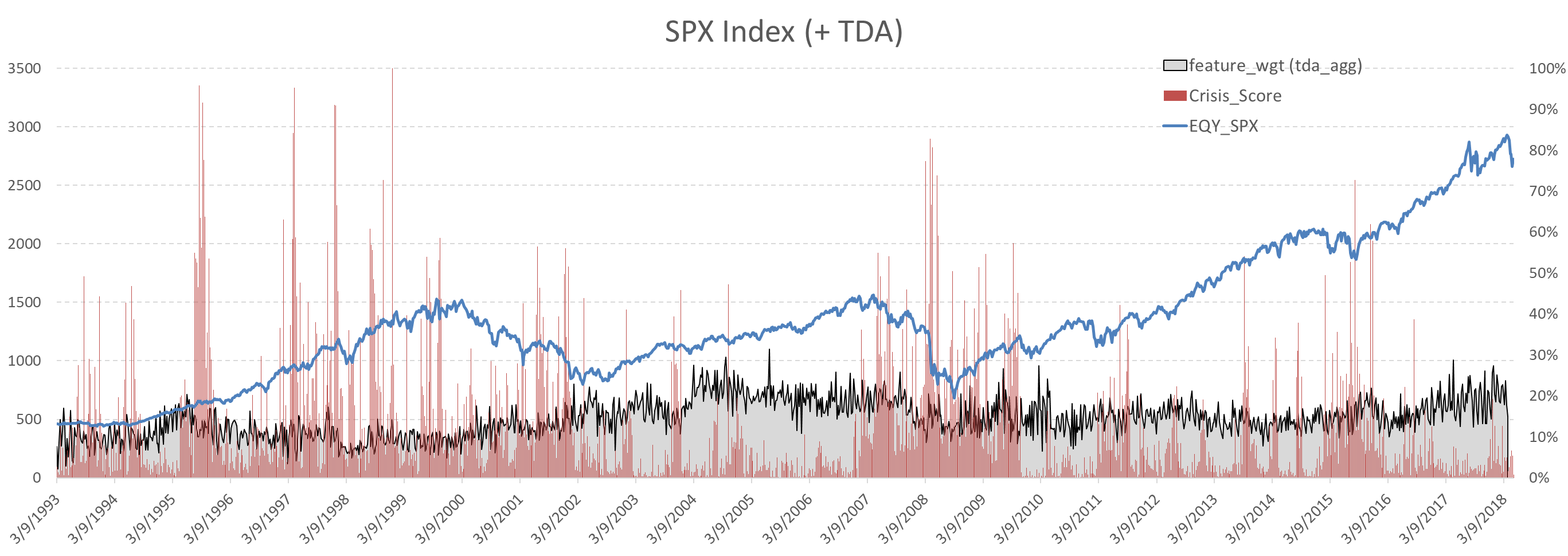
For FTSE 100 Index:

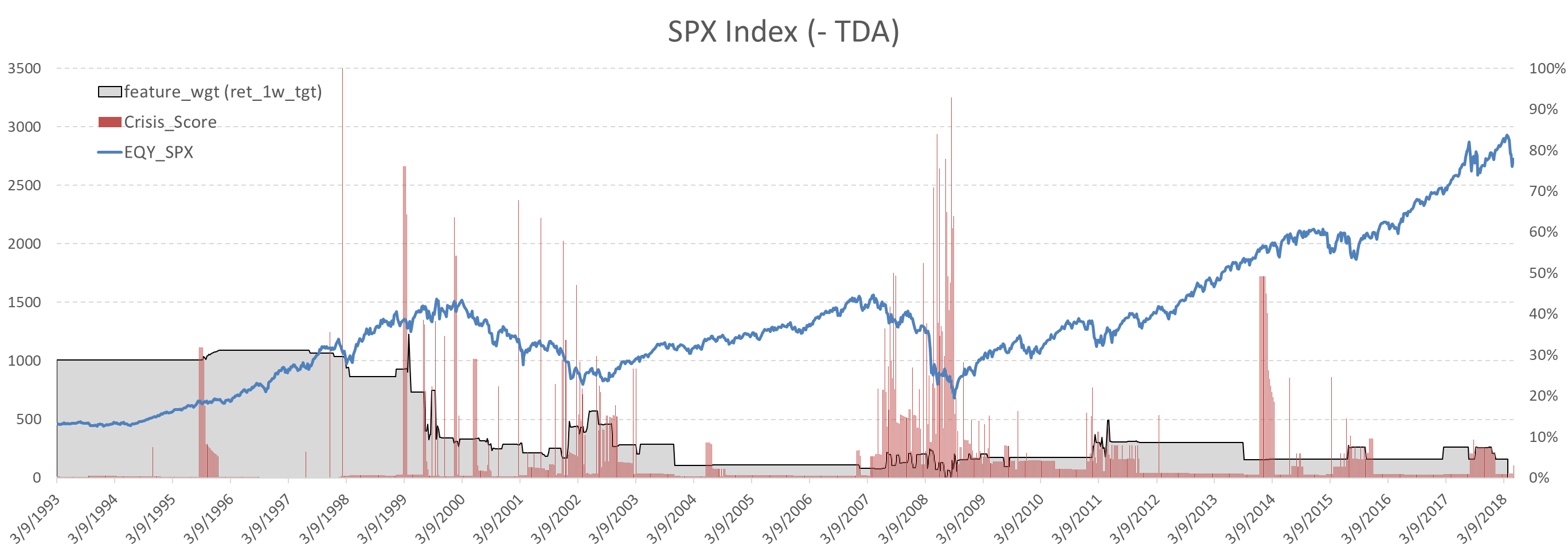


For a given confidence level (a chosen cutoff point in terms of probability, so that a prediction is positive if its probability is higher than this cutoff probability and vice versa), F1 score measures how good a classifier is by taking into account both the precision and recall ratios, i.e. in our case, “*how many positive predictions are correct out of all the positive predictions*” and “*how many positive predictions are correct out of all the actual crashes*”, respectively. Therefore, by design, F1 score punishes both false negatives and false positives: taking the S&P 500 backtesting for example, when the cutoff is low, most of the predictions will be considered positive but the reality is the opposite, which results in large number of false positives and hence lower precision, but this could be compensated by higher recall, because a lot of the true crashes will be “indiscriminatingly” predicted correctly, and this lower false negatives will reflect in higher recall ratio; on the other hand, when the cutoff is high, most of the predictions might be considered negative and would miss out the true crashes, resulting in high number of false negatives and low recall ratio.

From Figure 5, all models trained without TDA perform poorly when the cutoffs, i.e. the convictions or probabilities of coming market crashes, are beyond the 50% mark, with the F1 scores unforgivingly dropping to near 0%; on the other hand, the models trained with TDA features sustain a better prediction at high cutoffs, with F1 score stabilised at 10%. Admittedly, there is still huge room to improve, but consider the noisiness of financial market data, these results of 10%’s and above are the fair reflection and summary of TDA models’ performance during the time when markets are really in crisis mode, as well as, in the same window, the increased model sensitivity when markets are not, i.e. punished by higher crash scores or probability during calm periods.

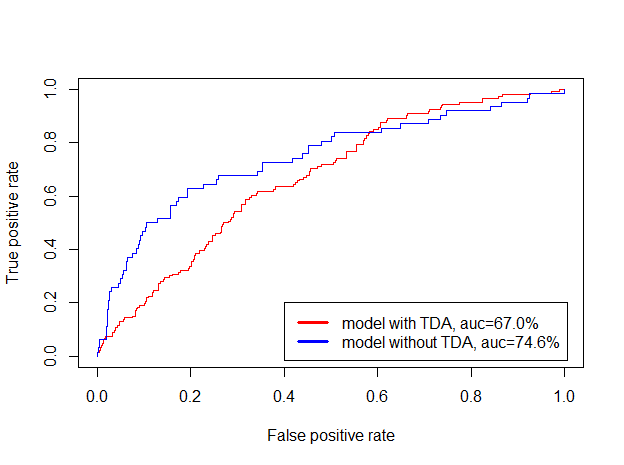
In the case of the S&P 500, we compare the crash probabilities of models with and without TDA features in Figure 6 and we see that crash probabilities are in general higher for TDA model than for non-TDA model, regardless of market regimes.





As a result, the over performance measured by Area Under Curve for TDA models are slightly lower than for those without (also in Figure 5), which can be seen as a “cost” in order to provide more robust crash signals during crisis periods. The Area Under Curve measure for both models are broadly of the same magnitude as for the various models in the study by Chatzis et al (2018).

For S&P 500 Index:



Same pictures will be seen for the other indices which we leave to the appendix.