Discussion

Model diagnosis

Having seen the model performance in Table~\ref{tab:res}, it is natural to ask why certain models perform better than others. While logistic regression models are expected to perform less well than more complex models, it is worthwhile to explore why the RF-RW model and the RNN-BiLSTM-focal model do not perform as well as the RF-CV model.

After further investigation of the model performance on the training data, I find that there is an overfitting problem for the RF-RW model. The model has a 99.5\% balanced accuracy rate on the training data but only 75\% on the testing data. The model performs worse on the testing data than on the training data. On the contrary, the RF-CV model does not see this problem.

Why would the RF-RW overfit but not the RF-CV? It might be that the re-weighting method makes the model over-trained for the bubble cases in the training data. The re-weighting method replicates the bubble cases, which could cause the model to fit too much on peculiar bubble patterns in the training data. The patterns are not universal for all bubble cases, and they are likely not to be presented in the testing data. This is perhaps why the RF-RW model performs worse on the testing data. On the other hand, the RF-CV alters the decision threshold. This method makes the model focuses on the bubble cases generally, but not restricted to cases seen in the training data. Without the over-training on the peculiar bubble patterns in the training data, the RF-CV model avoids the overfitting problem.

For the RNN-BiLSTM-focal model, the prediction is no better than a naïve predictor. I have tried various techniques including adjusting hypermeters, e.g., the number of neurons, changing layers, e.g., using LSTM rather than BiLSTM, and changing the loss function, e.g., changing the hypermeters of the focal function. None of these techniques works. This is quite a disappointment given the relevant research that endorses the model.

I examine the probability output of the model. Its distribution is shown in Figure~\ref{fig:rnn}. It shows that the variance of probability outputs of the model is small: almost all outputs are around 0.37. Also, the outputs of bubbles and non-bubbles overlap with each other. The model cannot distinguish the two based on the output. Based on this output, I suspect that the model is under-trained. It is likely because the model is too complex for our data size. A study on the data size impact on LSTM efficiency finds that three-year daily data is sufficient for training the model. \parencite{datasize} That is approximately 1000 data points. However, in this study, we only have 556 data points to train the RNN. It could be that the training data is insufficient to train this model.

It is worth noting that the above model diagnoses for the RF-RW and the RNN-BiLSTM-focal are largely speculative. Due to the complexities of the models, there does not seem to be an easy way to check my explanations.

Robustness analysis

I did the robustness analysis for the best performing model, the RF-CV model. As discussed in the data section, the quantitative definitions of market crashes and bubbles are rather arbitrary. To check the model performance under different definitions, I alter the “one percentile” and “six months” part of my definition. The results are shown in Table~\ref{tab:rob}.

We see the model performance is reasonably robust. The model has balance accuracy rates around 95\% if the percentile in the market crash definition is lower than 1 and the number of periods in the bubble definition is between 5 and 12. However, the model performs considerably worse for a short-term bubble definition (3-month period) and a less extreme definition of market crashes (5 percentile). This suggests that our model is limited in predicting extreme market crashes in a relatively long timeframe.

The result is not unexpected. The most important factors of the model are market fundamental indicators and long-term trends, so it is plausible that the model based on these factors has less predictive power in the short term. On the other hand, the CV-turned decision threshold method is adopted specifically for the imbalanced data problem. If the data is less imbalanced due to a less extreme market crash definition, the method will not be effective.

Limitations

Besides the short-term and less extreme limitation as discussed in the robustness analysis subsection, two limitations are worth mentioning.

The first limitation is regarding the interpretation of the results. A balanced accuracy rate should not be confused with a precision rate. A balanced accuracy rate stands for the balanced percentage of correctly identified cases. On the other hand, a precision rate means among all cases that are predicted as bubbles, the percentage of cases that are actual bubbles. While the RF-CV model has a high balanced accuracy rate, the precision rate is much lower, at 36.7\%. This means that if the model gives a “bubble” prediction, there is a 36.7\% possibility that a market crash would in the next six months. From a practical perspective, the user of the model should heed the possibility of false alarms when the model predicts bubbles.

Why would the two metrics differ? It can be best illustrated by the confusion matrix of the model prediction. As shown in Table~\ref{rf\_cfm}, the model correctly identifies most of the cases in both categories, but precision is modest. Although the portion of misclassified non-bubble cases is low, the number of false positives is high. Non-bubble cases are much more than bubble cases so that a small portion of misclassified non-bubbles are comparable to the correctly identified bubbles. This drives the precision much lower than the accuracy.

The second limitation is that the training data and testing data are not independent. While they are split randomly from the full data, they are both from a similar time period. For example, the data of May 2008 is in the training set, whereas the data of June 2008 could be in the testing set. The two periods are positively correlated due to their proximity in time. At the training stage, the model fitter learns that May 2008 is in a situation of bubble and incorporates its pattern into the model. When faced with the data of June 2008 in the testing set, it is easy for the model to correctly classify it as a bubble case since it has seen a similar instance in the training data. However, for the real-world application, the model would encounter cases that have much less similarity to the training data. It is dubious if the model can perform would perform equally well. In short, the dependence between the training data and the testing data makes the testing performance of the models overstated.