PREDICTING JUMPS IN US HIGH YIELD SPREADS

March 2018

Sonja Tilly

Approach and method

- Project
- Dataset
- Feature engineering
- Target variable
- Classification
  - Building model
  - Feature elimination
- Results
  - Feature importances
  - Partial dependencies
- Conclusion

Project

- Predict jump risk in high yield spreads
- “Jump” is arbitrarily defined as increase in OAS of more than 30bps in 30 days using the Bloomberg Barclays US Corporate High Yield Average OAS (LF98OAS Index)
- This is a classification task

Dataset

- The data contains rows of economic indices in daily observations since 1997, showing 7,364 rows and 33 columns.

```
data.columns
```

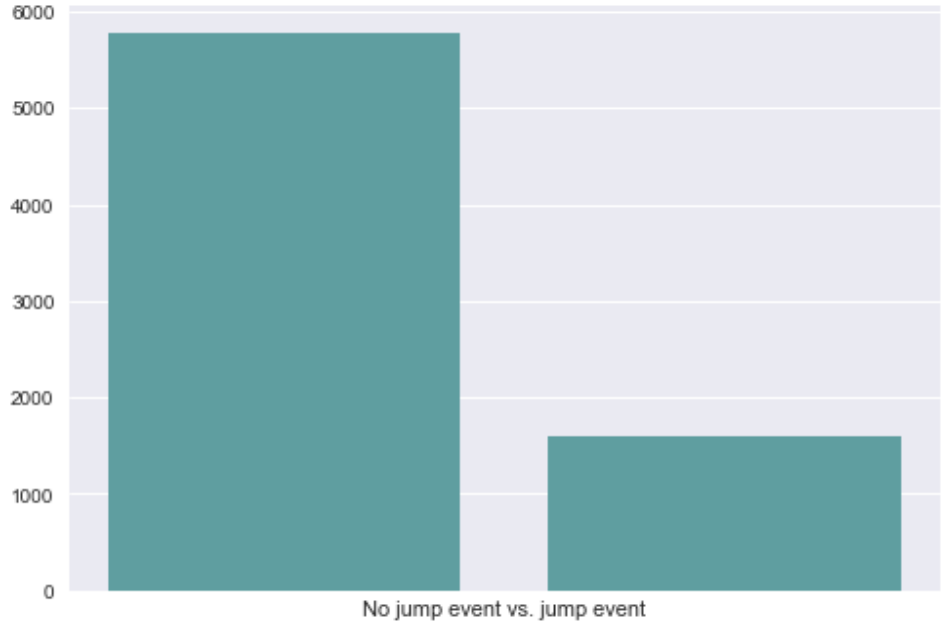
```
Index([u'FDTR Index', u'VIX Index', u'USGG5YR Index', u'USYC2Y10 Index',
       u'CPI YOY Index', u'USURTOT Index', u'OUTFGAF Index', u'CLA Comdty',
       u'CONSSENT Index', u'SCGRRAI Index', u'EMB US Equity',
       u'SPE AUTO Index', u'SPE CARD Index', u'LEI CHNG Index',
       u'GFSIFFND Index', u'XLF US Equity', u'XLE US Equity', u'XLK US Equity',
       u'XLV US Equity', u'XLI US Equity', u'XLY US Equity', u'XLB US Equity',
       u'PCUSEQTR Index', u'GFSIRLIQ Index', u'CIGMGRAM Index', u'SPX Index',
       u'US0003M Index', u'AHE YOY% Index', u'NHSPSTOT Index',
       u'NAPMNEWO Index', u'CPTICHNG Index', u'USGGBE05 Index',
       u'LF98OAS Index'],
      dtype='object')
```

Feature engineering

- Address time lags in indicators.
- Time windows of 30, 90, 365, 1,095 and 1,825 days have been defined.
- Data normalisation. The z-score has been calculated for all features for the above defined time windows. For features with monthly data, 30 and 90 day windows have been omitted.
- Level change or percentage (depending on convention) change have been calculated for all features for the above defined time windows.
- Drop the original input indices.
- Drop any feature derived from LF98OAS Index.
- Replace infinite and nan values with zero.

Target variable

- Transform target variable into binary numbers for classification (“Signal”):
- LF98OAS Index change in 30 days is larger or equal to 30 bps: 1, else: 0
- Apply a negative lag of 5 days to target variable i.e. the model will predict jump risk 5 days into the future.
- The target variable is unbalanced, with no-jump events being c 3.6 times more common than jump events within a time frame of 30 days.



Classification

- Define predictor and target variables
- Split out training and testing sets
  - Training: first 5,301 observations
  - Testing: last 2,063 observations
- Build Gradient Boosted Tree classification model
- Model evaluation metric: log loss
- Evaluate model performance using cross-validation with forward-chaining.

Classification

- Tuning the following parameters had positive impact:
  - Account for unbalanced classes (sample_weight);
  - Increase number of boosting stages to perform (n_estimators);
  - Reduce the number of features considered when looking for the best split (max_features).

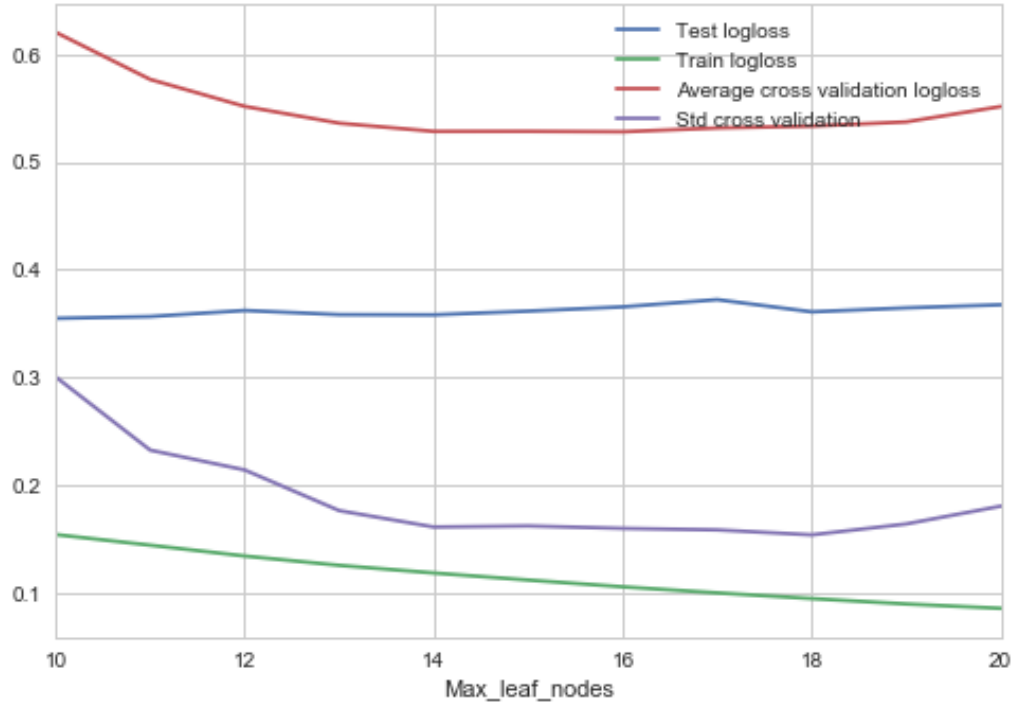
Feature elimination

- Apply recursive feature elimination with cross-validation.
- The optimal number of features determined with this technique is 9.
- Addressed collinearity by calculating the variance inflation factor (vif) for each feature and removing those with a vif over 5.
- Remaining number of features: 9

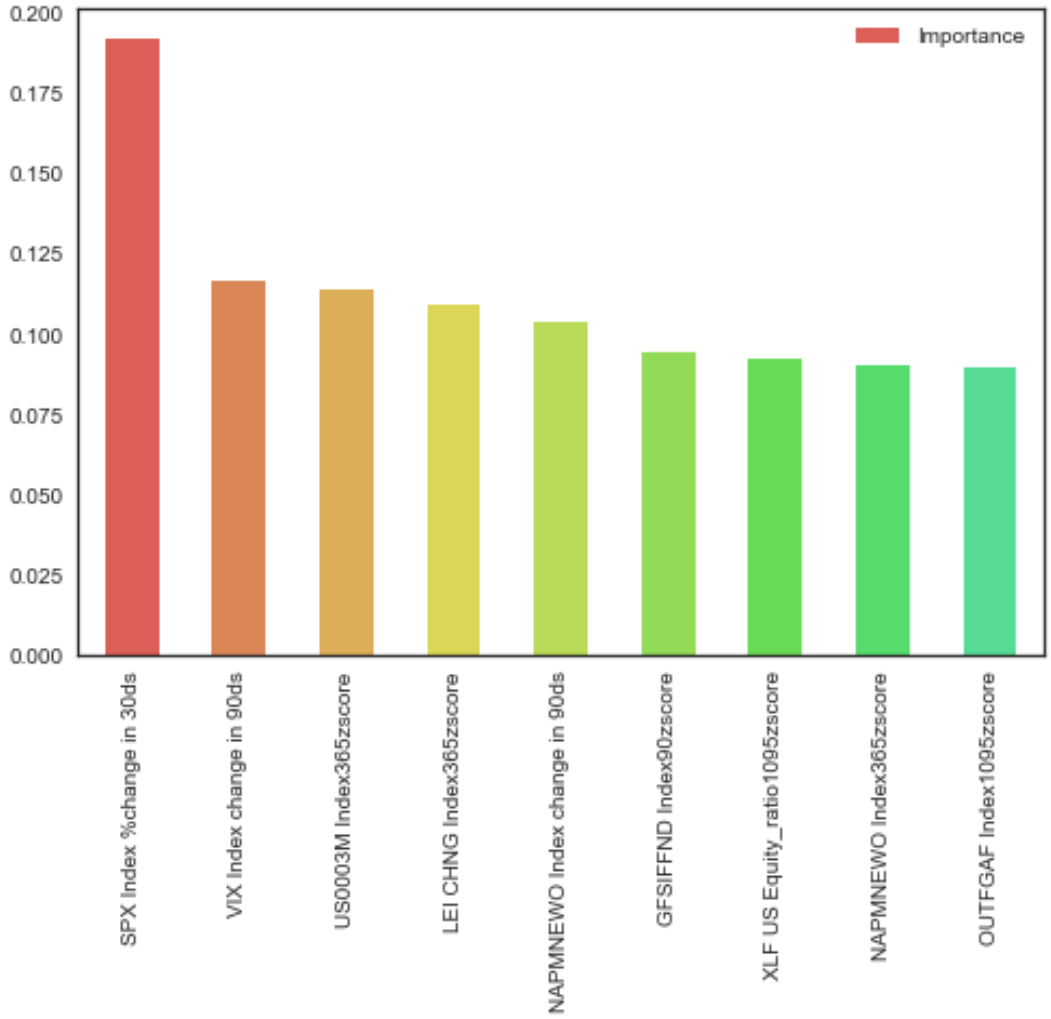
Results: model performance

Before feature elimination

- Log loss score on training set: 0.0857.
  This equates to a probable model accuracy of 91.79%.
- Log loss score on testing set: 0.3674.
  This equates to a probable model accuracy of c 70%.



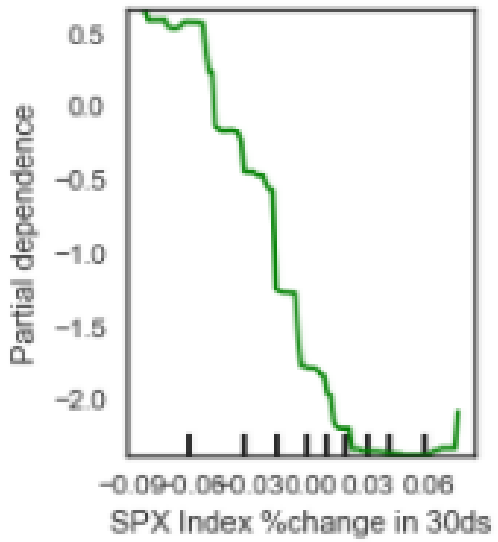
Results: feature importances



Top 10 features:

1. S&P500 % change in 30 days
2. VIX index change over 90 days
3. 3M Libor to Overnight Indexed Swap Spread z-score over 365 days
4. Conference Board Leading indicator index change of 365 days
5. ISM New Orders index change in 90 days
6. BofA ML Fund Flow index z-score over 90 days
7. US Financials to S&P500 ratio z-score over 1095 days
8. ISM New Orders index change in 365 days
9. Philadelphia Fed Business Outlook Survey index z-score over 1095 days

Results: partial dependencies

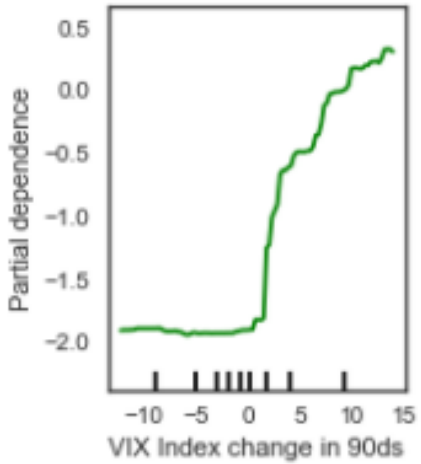


Feature importance rank: 1

S&P500 % change in 30 days is the most important feature in this model.

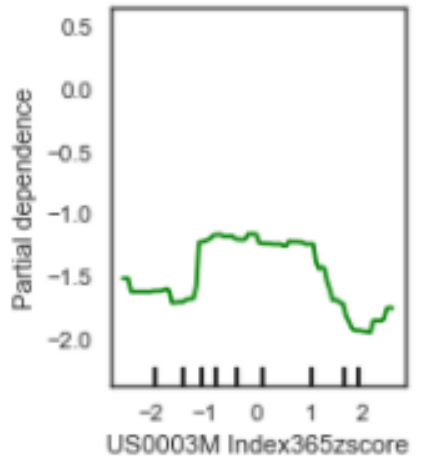
Its contribution to log loss decreases the more the feature moves into positive territory, then ticks up again at around .

Results: partial dependencies



Feature importance rank: 2

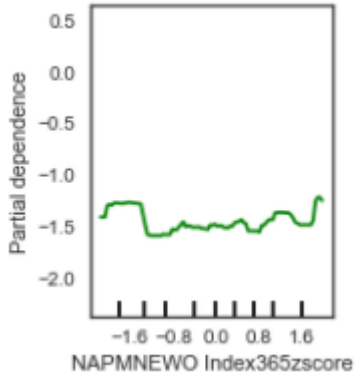
Importance steeply increases as the VIX index change over 90 days moves into positive territory.



Feature importance rank: 3

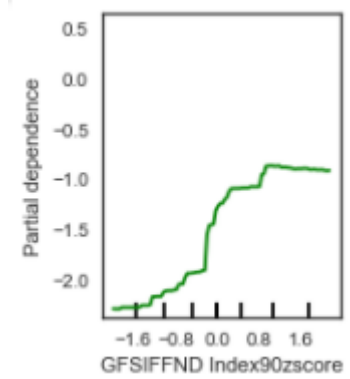
Importance is highest when the 3M Libor to Overnight Indexed Swap Spread z-score over 365 days ranges between -1 and +1.

Results: partial dependencies



Feature importance rank: 4

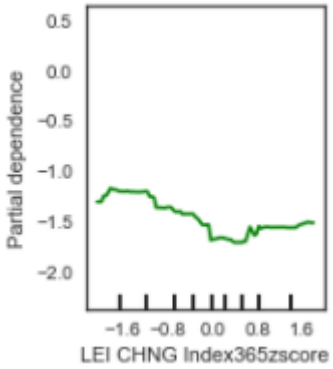
Importance of Conference Board Leading indicator index change of 365 days fluctuates in a tight range.



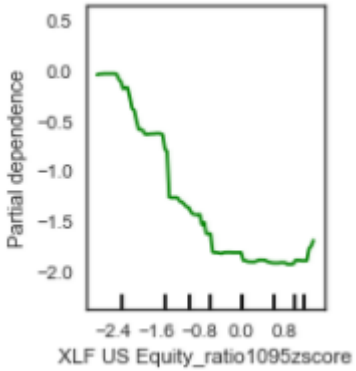
Feature importance rank: 5

Importance of the ISM New Orders index change in 90 days steeply increases before plateauing at around 1.

Results: partial dependencies

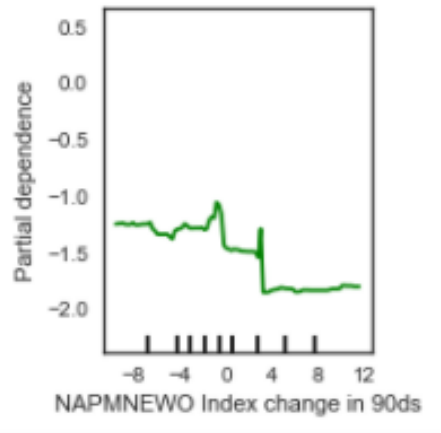


Feature importance rank: 6
The importance of BofA ML Fund Flow index z-score over 90 days slightly decreases as the feature moves towards zero and modestly picks up as the features moves more positive.



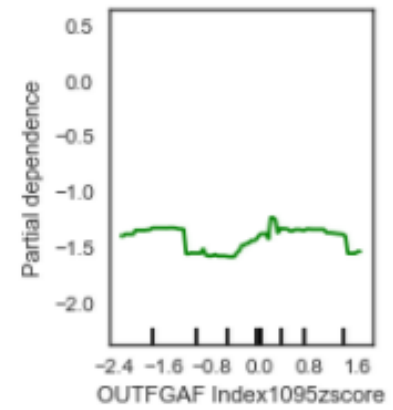
Feature importance rank: 7
The importance of US Financials to S&P500 ratio z-score over 1095 days increases as the feature moves into positive territory.

Results: partial dependencies



Feature importance rank: 8

The importance of the ISM New Orders index change in 365 days is declining as the features move into positive territory.



Feature importance rank: 9

The importance of the Philadelphia Fed Business Outlook Survey index z-score over 1095 days oscillates in a relatively tight range.

Benchmark model

- A logistic regression model was built as benchmark model.
- Unbalanced classes have been taken into consideration.

- Log loss score on training set: 0.3674.
- Log loss score on testing set: 0.8327.
- This translates into a probability of correct model forecasts of 69.25% and 43.49%, respectively.

Conclusion

- The Gradient Boosting Tree classifier outperforms the benchmark model by a comfortable margin.
- The key drivers of jump risk in US high yield spreads consist of equity market risk, volatility, market risk aversion, fund flows, the degree of credit risk within the banking sector.
- This seems intuitive and is in line with other research.

Criticism

- While the model outperforms the benchmark model, the log loss remains substantial and further efforts could be made to improve performance, such as more feature engineering and experimentation with other algorithms, e.g. deep neural networks.

Appendix I

Log Loss

Log Loss quantifies the accuracy of a classifier by penalising false classifications. Minimising the Log Loss is basically equivalent to maximising the accuracy of the classifier, but there is a subtle twist which we'll get to in a moment.

In order to calculate Log Loss the classifier must assign a probability to each class rather than simply yielding the most likely class. Mathematically Log Loss is defined as

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\, p_{ij}$$

where N is the number of samples or instances, M is the number of possible labels, yijyij is a binary indicator of whether or not label j is the correct classification for instance i, and pijpij is the model probability of assigning label j to instance i. A perfect classifier would have a Log Loss of precisely zero. Less ideal classifiers have progressively larger values of Log Loss. If there are only two classes then the expression above simplifies to

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i\log\, p_i + (1-y_i)\log\,(1-p_i)].$$

Log Loss heavily penalises classifiers that are confident about an incorrect classification. For example, if for a particular observation, the classifier assigns a very small probability to the correct class then the corresponding contribution to the Log Loss will be very large indeed. This is going to have a significant impact on the overall Log Loss for the classifier.

Source: http://www.exegetic.biz/blog/2015/12/making-sense-logarithmic-loss/

Appendix II

Variance inflation factor

A variance inflation factor detects multicollinearity. Multicollinearity exists when there is correlation between predictors (i.e. independent variables) in a model. Its presence can adversely affect model performance. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

R-squared is derived from taking on independent variable and regressing it against every other predictor in the model.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

Source: http://www.statisticshowto.com/variance-inflation-factor/