Dark Pools Stock Trading Prediction Model

Liyu Ma[1], Z. Yao[1], S. Moazeni[1], and R. Collado[1]

[1]School of Business at Stevens Institute of Technology

Introduction and Motivation

Dark pools are private venues where investors can exchange large amounts of stock without overly impacting market price.

Defining features:

- Little transparency of trade execution and asymmetric information flow.
- Trades are executed within the spread.
- Multiple dark pools venues form a fragmented market with dozens of venues in the United States alone.

Motivation:

- Dark pools are ubiquitous in modern trading.
- Due to its asymmetric flow of information, dark pools are a challenge to model and master.
- Current finance literature shows limited understanding of the inner workings of dark pools.
- Current trends in dark pools trading follow a purely algorithmic trading approach where vast amounts of asymmetric data is collected.

Goals and Challenges:

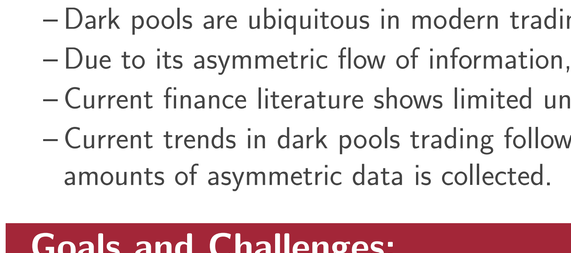
Goals: To develop machine learning classification methods capable of exploiting the asymmetric data to predict liquidity in dark pool trading

Challenges:

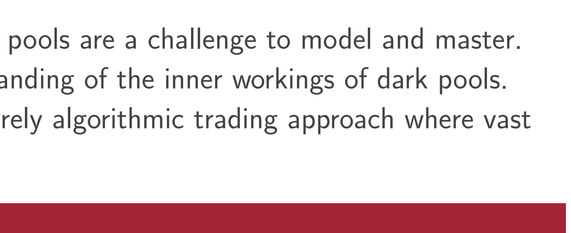
- Asymmetric data containing only one-sided information of orders and trades.
- Highly imbalanced data set (less than 1% orders become trades) creates serious technical difficulties for standard classification methods.
- Big data set that require attention to our ML optimization methods.

Data Composition:

- 20 million data points for 1 month data (June 2017).
- Focus on **UBSA** venue and order parameters **TIF:** DAY and **PegInst:** None.
- Generated features based on historical data: **L1**, **L3**, and **L5**.
- Additional features: **StartTime**, **Symbol**, **Venue**, **VenueType**, textbfSecurityCategory, **Sector**, **MktCap**, and **Adv20d**.

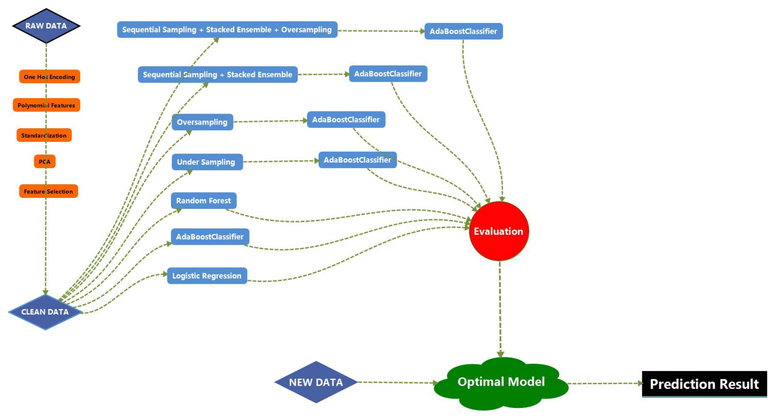


(a) Orders of symbol 22227111 placed on all dark venues over 1 trading day.



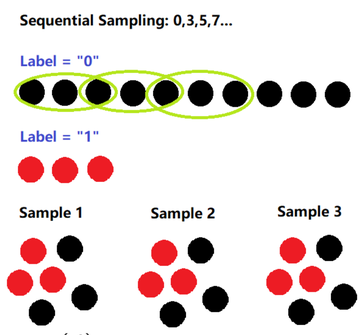
(b) Orders of symbol 3535 placed on CAES venue over 1 trading day.

Machine Learning Model:

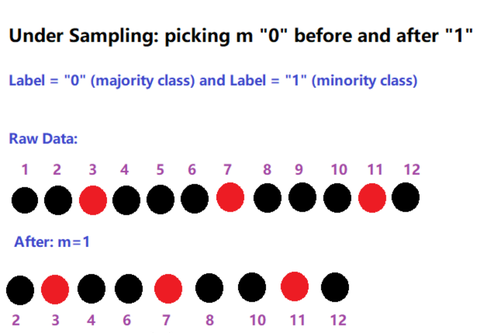


First, multiple methods are applied to perform data cleaning and features selection. Then, we apply a battery of ML methods with differing techniques to the clean data. Finally, all methods are ranked and the best method is used on predictions.

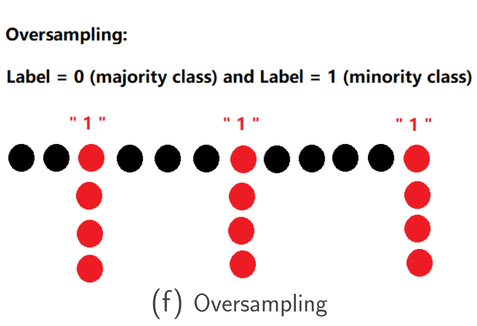
Sampling Methods:



(d) Sequential Sampling



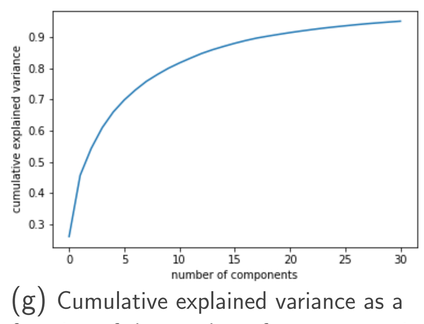
(e) Undersampling



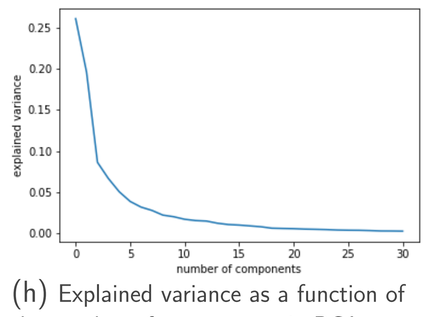
(f) Oversampling

PCA & Feature Selection:

PCA Results:



(g) Cumulative explained variance as a function of the number of components in PCA.

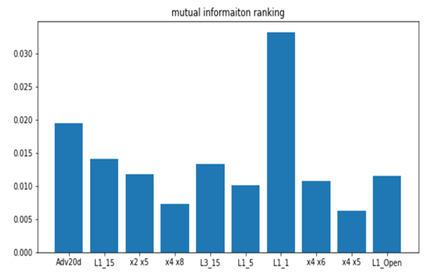


(h) Explained variance as a function of the number of components in PCA.

Mutual Information Coefficient & Feature Selection:

```
*** mRMR features ***
Order    Fea     Name    Score
1        6       MktCap_cat      0.016
2        4       SecurityCategory_cat    0.001
3        5       Sector_cat      0.000
4        1       Symbol_cat      0.000
5        2       VenueType_cat   0.000
6        3       Side cat        0.000
```

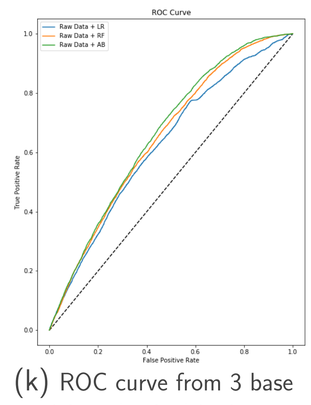
(i) Categorical feature ranking based on output from maximum–relevance minimum–redundancy algorithm.



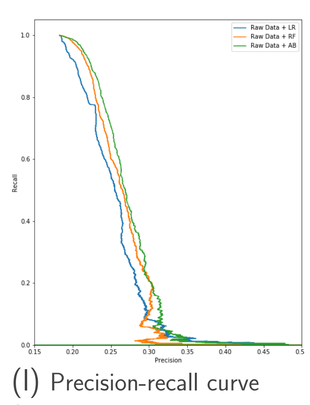
(j) Top 10 continuous features ranking based on output from multivariate mutual information algorithm.

Results:

Base Model: Logistic Regression, Random Forest, and AdaBoost

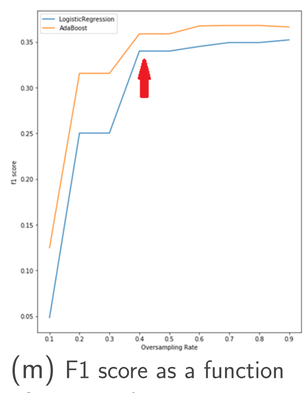


(k) ROC curve from 3 base models trained with clean data.

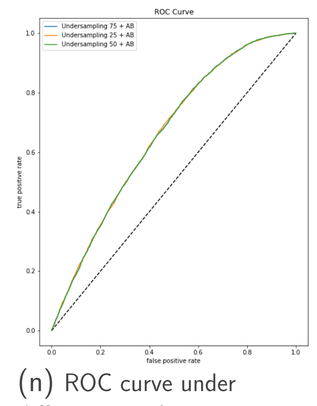


(l) Precision-recall curve from 3 base models trained with clean data.

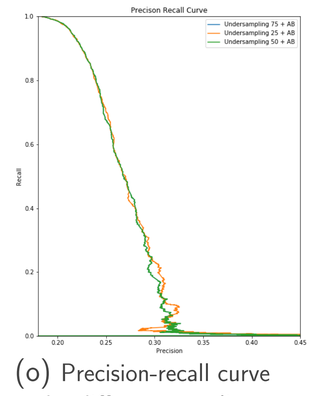
Cross Validation:



(m) F1 score as a function of oversampling rate.

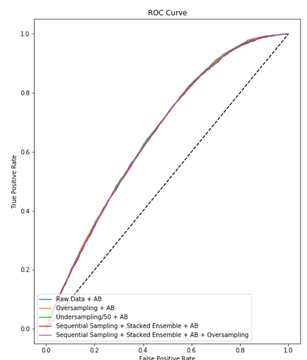


(n) ROC curve under different sampling rates.

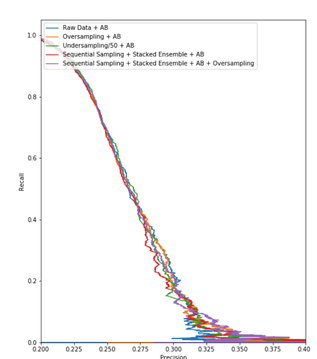


(o) Precision-recall curve under different sampling rates.

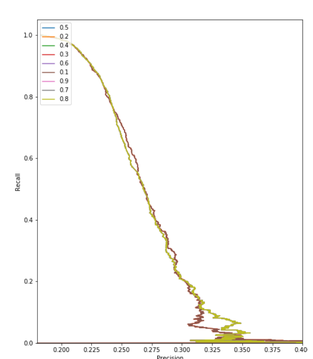
Model Optimization:



(p) ROC curve from AdaBoost models trained by different techniques.



(q) Precision-recall curve from AdaBoost models trained different techniques.



(r) Precision-recall curve from models trained on different oversampling rates.

Conclusions:

- In order to deal with asymmetric and highly imbalanced data set we employed several advanced machine learning techniques to improve the classification of the minority class.
- Comparison of our results against vanilla out-of-the-box machine learning methods shows an increase in classification performance.
- Highly imbalanced data set such as the dark pools trading data tend to exhibit limits on classification. Our analysis give us strong indication that we are reaching such limits within the selected framework.
- Our results show AdaboostClassifier performing better than the other studied algorithms.
- By studying the multivariate mutual information between features and output variable, it is clear that the existing features share limited information with output variable. Hence, generating useful features or adding more features into the model are reasonable directions of future work.

https://www.stevens.edu/school-business lma18@stevens.edu