Karlsruhe Institute of Technology

Institute for Finance

Department of Financial Engineering and Derivatives
Prof. Dr. Marliese Uhrig-Homburg

Master Thesis

Option Trade Classification using Machine Learning

Author: Markus Bilz
Mathystr. 14-16 // XI-11
76133 Karlsruhe
E-Mail: markus.bilz@student.kit.edu

Karlsruhe, March 31, 2023

Contents

List of Figures

List of Tables

1 Introduction (2 p)

2 Related Work (3 p)

3 Rule-Based Approaches (5.75 p)

3.1 Basic Rules (2.5 p)

3.1.1 Quote-Rule (0.5 p)

Algorithm 1: `quote`

```
Input  : $t_i$ trade price at $i$, $a_i$ ask price at $i$, and $b_i$ bid price at $i$.
Output: $o_i \in \{-1, 1\}$ trade initiator for $i$-th trade.
$m_i \leftarrow \frac{1}{2}(a_i + b_i)$                    /* mid spread at $i$ */
if $t_i > m_i$ then
    return $o_i = 1$
else if $t_i < m_i$ then
    return $o_i = -1$
else
    return
```

3.1.2 Tick Test (1 p)

3.1.3 Depth Rule (0.5 p)

Grauer et al. (2022) promote an alternative to improve the classification performance of midspread trades. In their *depth rule*, they infer the trade initiator from the depth of the ask and bid. Based on the observation that an exceeding bid or ask size relates to higher liquidity on one side, trades are classified as buyer-initiated for a larger ask size and seller-initiated for a higher bid size.

As shown in Algorithm 2, the depth rule classifies midspread trades only, if the ask size differs from the bid size, as the ratio between the ask and bid size is the sole criterion for assigning the initiator. To sign the remaining trades, other rules must be employed thereafter.

Algorithm 2: `depth`

```
Input  : $t_i$ trade price at $i$, $a_i$ ask price at $i$, $b_i$ bid price at $i$, $\tilde{a}_i$ ask size at
         $i$, and $\tilde{b}_i$ bid size at $i$.
Output: $o_i \in \{-1, 1\}$ trade initiator for $i$-th trade.
$m_i \leftarrow \frac{1}{2}(a_i + b_i)$                              /* mid spread at $i$ */
if $t_i = m_i$ then
    if $\tilde{a}_i > \tilde{b}_i$ then
        return $o_i = 1$
    else if $\tilde{a}_i < \tilde{b}_i$ then
        return $o_i = -1$
    else
        return
else
    return                                                 /* apply secondary rule */
```

In a similar vein, the *trade size rule* reuses the ask and bid quote size to improve the classification performance of trades where the trade size equals the ask quote or bid quote sizes.

3.1.4 Trade Size Rule (0.5 p)

Algorithm 3: `tradesize`(t_i, a_i, b_i)

```
Input  : $\tilde{t}_i$ trade size at $i$, $\tilde{a}_i$ ask size at $i$, and $\tilde{b}_i$ bid size at $i$.
Output: $o_i \in \{-1, 1\}$ trade initiator for $i$-th trade.
if $\tilde{a}_i = \tilde{t}_i$ and $\tilde{b}_i \neq \tilde{t}_i$ then
    return $o_i = -1$
else if $\tilde{b}_i = \tilde{t}_i$ and $\tilde{a}_i \neq \tilde{t}_i$ then
    return $o_i = 1$
else
    return                                                 /* apply secondary rule */
```

Algorithm 4: `lee-ready` (t_i, a_i, b_i)

```
Input  : $t_i$ trade price at $i$, $a_i$ ask price at $i$, and $b_i$ bid price at $i$.
Output: $o_i \in \{-1, 1\}$ trade initiator at $i$.
$m_i \leftarrow \frac{1}{2}(a_i + b_i)$                                  /* mid spread at $i$ */
for $1, \cdots, I$ do
    if $t_i > m_i$ then
        return $o_i = 1$
    else if $t_i < m_i$ then
        return $o_i = -1$
    else
        return $o_i = \texttt{tick}(t_i, a_i, b_i)$                      /* see above */
    end
end
```

Algorithm 5: `emo`

```
Input  : $t_i$ trade price at $i$, $a_i$ ask price at $i$, and $b_i$ bid price at $i$.
Output: $o_i \in \{-1, 1\}$ trade initiator at $i$.
for $1, \cdots, I$ do
    if $t_i = a_i$ then
        return $o_i = 1$
    else if $t_i = b_i$ then
        return $o_i = -1$
    else
        return $o_i = \texttt{tick}(t_i, a_i, b_i)$                      /* see above */
    end
end
```

3.2 Hybrid Rules (3.25 p)

3.2.1 Lee and Ready Algorithm (1 p)

3.2.2 Ellis-Michaely-O'Hara Rule (0.5 p)

3.2.3 Chakrabarty-Li-Nguyen-Van-Ness Method (0.5 p)

3.2.4 Rosenthal's Rule (0.75 p)

4 Supervised Approaches (12 p)

4.1 Selection of Approaches (2 p)

4.2 Gradient Boosted Trees (2 p)

4.2.1 Decision Tree (0.5 p)

4.2.2 Gradient Boosting Procedure (1 p)

4.2.3 Adaptions for Probabilistic Classification (0.5 p)

4.3 Transformer Networks (8 p)

4.3.1 Network Architecture (2.5 p)

4.3.2 Attention (0.5 p)

4.3.3 Positional Encoding (0.5 p)

4.3.4 Embeddings (0.5 p)

4.3.5 Extensions in TabNet (2 p)

4.3.6 Extensions in TabTransformer (2 p)

5 Semi-Supervised Approaches (8 p)

5.1 Selection of Approaches (2 p)

5.2 Extensions to Gradient Boosted Trees (2 p)

5.3 Extensions to TabNet (2 p)

5.4 Extensions to TabTransformer (2 p)

6 Empirical Study (19.5 p)

6.1 Environment (0.5 p)

6.2 Data and Data Preparation (6 p)

6.2.1 ISE Data Set (0.5 p)

6.2.2 CBOE Data Set (0.5 p)

6.2.3 Generation of True Labels (0.5 p)

6.2.4 Feature Engineering (4 p)

6.2.5 Train-Test Split (0.5 p)

6.3 Training and Tuning (10 p)

6.3.1 Training of Supervised Models (4 p)

6.3.2 Training of Semi-Supervised Models (4 p)

6.3.3 Hyperparameter Tuning (2 p)

6.4 Evaluation (3 p)

6.4.1 Feature Importance Measure (2 p)

6.4.2 Evaluation Metric (1 p)

7 Results (12 p)

7.1 Results of Supervised Models (3 p)

7.2 Results of Semi-Supervised Models (3 p)

7.3 Feature Importance (3 p)

7.4 Robustness Checks (3 p)

8 Discussion (3 p)

9 Conclusion (2 p)

10 Outlook (0.5 p=67.75 p)

References

Grauer, Caroline, Philipp Schuster, and Marliese Uhrig-Homburg (2022). "Option Trade Classification". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: `10.2139/ssrn.4098475`.