IIROC OCRCVM
Investment Industry Regulatory Organization of Canada
Organisme canadien de réglementation du commerce des valeurs mobilières

Identifying Trading Groups
Methodology and Results

Trading Review and Analysis
Investment Industry Regulatory Organization of Canada

September 9, 2014

Acknowledgements:

Baiju Devani, Director Analytics, IIROC
Ad Tayal, Lead Data Scientist, IIROC
Lisa Anderson, Team Lead Analytics, IIROC
Dawei Zhou, Sr. Programmer Analyst, IIROC
Juan Gomez, Trading Analyst, IIROC
Graham W. Taylor[1,2]

[1] IIROC Consultant
[2] School of Engineering, University of Guelph, Guelph, ON, gwtaylor@uoguelph.ca

Table of Contents

1. Introduction

In the last decade, financial markets and equity markets in particular have become increasingly complex. This complexity is driven by technological advances, increased competition and the arrival of new types of market participants. In this increasingly technology-based marketplace, it has become critical to identify distinct groups of participants based on their trading footprints and look at the interactions and impact of these groups on key market structure issues such as market quality, fairness and integrity.

This report outlines a novel and robust methodology for classifying distinct user groups using a supervised machine learning method. Previous work by IIROC had focused on either one dimension such as order-to-trade ratio or a select few dimensions [1, 2]. This approach builds upon our previous work but is significantly different in two ways:

1. We apply supervised learning through the use of a support vector machine (SVM) method for the classification of user groups. The use of such machine learning approaches is well established in other domains and our results demonstrate its effectiveness in this context.
2. We utilize the richness of the data available to us and construct a set of 200+ features that characterize the behavior of each user. Our experiments indicate that use of a large feature set combined with robust methodology improves classification results over the use of a few hand-picked features by domain experts.

In the following sections we outline our methodology in detail and provide some results that evaluate the effectiveness of the classification algorithm. We then apply this classification scheme to segment the population of users over a study period and show some key interactions and metrics for each group.

2. Methodology

This section describes the methodology used to identify the type of activity associated with a trading UserID on Canadian equity marketplaces and discusses results for the period 3-Mar-2013 to 28-June-2013. This period corresponds with a period of stability with respect to policy or rule changes. Additionally, for this time period, IIROC had a self-reported group of retail UserIDs which could be used by our methodology.

We choose to categorize trading flows using the UserID field. The UserID, while not perfect, is the most useful field to consistently identify types of trading flows. Historically, the UserID was assigned by the marketplace to the broker for use by one individual trader. As the trading landscape has become more complicated and automated, use of the UserID has expanded and the order flow through individual UserIDs has therefore become more complex. The following cases illustrate this complexity:

- A single entity might have multiple UserIDs assigned by different marketplaces or brokers through which markets are accessed.

- A single UserID might be used for trading activity of different entities; for example, one UserID for all "order execution" retail flow.

Table 1 below illustrates the broad categories (or segments) of trading flow on the Canadian Equity markets, as differentiated by the account type (provided to IIROC in the regulatory feed) and the complexity of the strategies employed:

Table 1: Size and Complexity of Strategy by Account Type

	Simple/Small				Complex/Large/Intense	
Account Type	Client :	Retail	Day Trader	Institutional	Hedge Fund	Electronic Liquidity Provider
	Inventory:		Oddlot Market Maker	Client Facilitation		Broker Strategies
	Non-Client:	NC-Retail				
	Specialist :			Marketplace Specialist		
	Options Market Maker:			Options Market Maker		

We group trading activity into four broad categories:

1. High Frequency Trading (HFT)
    - Includes Electronic Liquidity Providers
    - May include Hedge Funds and Broker Strategies
2. Retail (RET)
    - Includes Retail and NC-Retail
3. Specialist (ST)
    - Includes Oddlot Market Maker and Marketplace Specialist
4. Sell side/Buy side (SB)
    - Includes Client Facilitation, Institutional, Broker Strategies
    - May include Day Trader, Hedge Fund, Options Market Maker

The goal is to find an automatic and objective rule that classifies each User ID into one of the four categories. We consider statistical machine learning methods to classify each UserID using a set of extracted features that measure the UserID's trading behavior. We manually label a small subset of UserIDs based on knowledge of the trading entity and use this labeled set to learn an inductive rule (i.e. supervised learning). Supervised machine learning methods, such as support vector machines and random forests, have proven widely successful in their ability to learn effective models even when the number of features is large with respect to the sample size and contain little or no assumptions on covariates [3, 4]. We follow a robust experimental methodology to train models and verify results.

i. Feature Extraction

We devise a comprehensive set of more than 200 features or characteristics based on the daily trading activity for each UserID. The features are designed to measure aggregate behavior of a UserID over a single trading day, encompassing the themes listed below. For each theme we have listed examples of the types of features which are input to the algorithm:

Table 2: Themes and Examples of Features

Theme	**Examples**
Trades and Orders	• Value of all trades • Number of amended orders • Order-to-trade ratio
Inventory Dynamics	• Percentage of trades marked SME[3] • Net Position
Speed Measures	• Order amendment speed • Percent of "simultaneous" orders
Account Type	• Percentage of trades with specific account type (i.e. Client) • Percentage of odd lot trades that traded with a specialist
Terms	• Percentage of orders using a "Good-Till" date • Percentage of Seek Dark Liquidity trades
Securities Traded	• Number of unique securities traded • Percent of trades by listing market
Crosses and Blocks	• Percent volume of crosses
Traded Market	• Percentage of trades by Market
Transactional Cost Management	• Percentage of active trades • Net rebates

[3] For more information on the SME (Short-Marking Exempt) order designation, see IIROC Notice 12-0300.

ii. Feature Smoothing

UserIDs may change behaviour day-to-day. Features are first calculated daily for each UserID. A moving one-month daily average is then calculated from the daily observations. The moving averages of the features are used by the machine learning models to represent each UserID. This smooths out day-to-day changes in the behavior of the UserID while still accommodating any long term changes in the fundamental uses of the UserID.

iii. Pre-Processing

Certain features exhibit extremely high skew and/or kurtosis, which implies that the feature takes on extreme values (i.e. the distribution of the feature has large tails). This can obscure patterns in the data. For example, Figure 1 shows the feature distribution of Order-to-Trade Ratio and Net Rebates. Extremely large (absolute) values of the feature overshadow variation in smaller (absolute) values.

To improve the distribution of a feature and make patterns more visible in the data, we apply the following transformation to features that have skew > 5 or kurtosis > 20[4]:

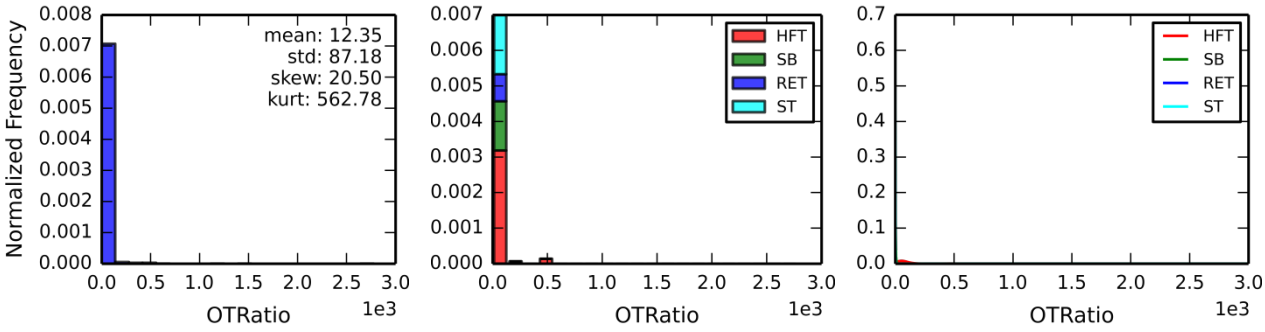
$$\tilde{x} = sign(x)\ln(1 + |x|),$$

where x is the original feature value and $\tilde{x}$ is the transformed feature value. The transformation can handle both positive and negative values. The effect of the log transformation is that small (absolute) values that are close together are spread further out and large (absolute) values that are spread out are brought closer together. For example, see Figure 2, which illustrates the log transformation of Order-to-Trade Ratio and Net Rebates. With the log transformation, patterns in the feature can be identified more clearly. This is particularly important for linear models which are unable to adapt to different feature representations automatically. Another byproduct of the pre-processing is making learning more numerically stable.

In the final step, we normalize each feature to mean zero and standard deviation one. Missing values are assigned a value of zero on the normalized scale, representing the mean value of the feature.

[4] In our experiments we found results to be insensitive to the exact cut-off values chosen for skew and kurtosis.

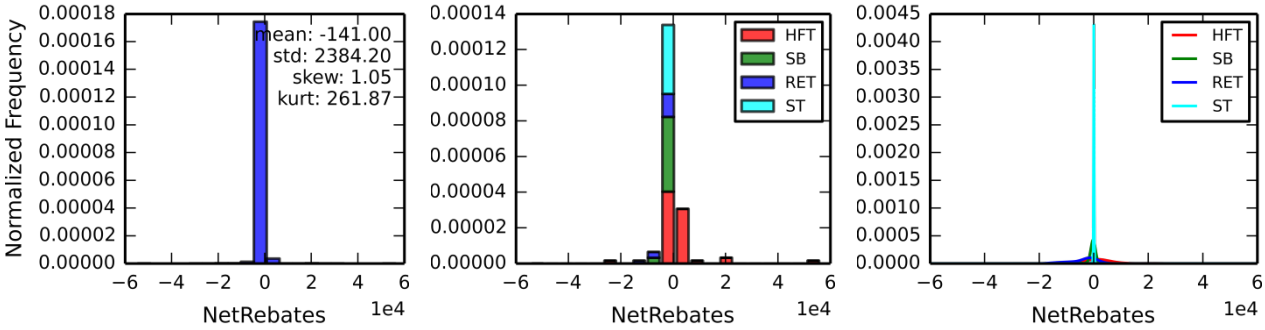
Figure 1: Distribution of Absolute Order-to-Trade Ratio and Net Rebates



(a) Histogram of all 2009 samples

(b) Histogram of 105 labeled samples

(c) Class conditional densities using labeled samples

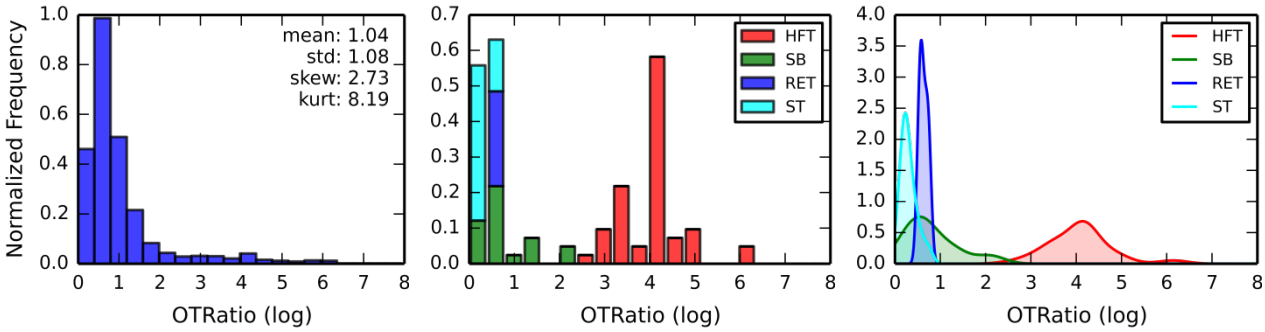


(d) Histogram of all 2671 samples

(e) Histogram of 113 labeled samples

(f) Class conditional densities using labeled samples

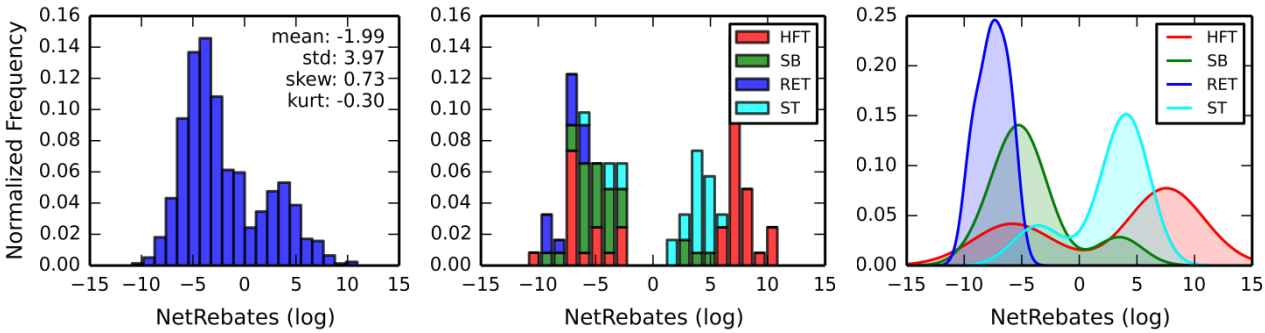
Figure 2: Distribution of Log Transformed Order-to-Trade Ratio and Net Rebates



(a) Histogram of all 2009 samples

(b) Histogram of 105 labeled samples

(c) Class conditional densities using labeled samples



(a) Histogram of all 2671 samples

(b) Histogram of 113 labeled samples

(c) Class conditional densities using labeled samples

iv. Labeled Training Set

We manually label a small set of UserIDs based on knowledge of the trading entity as of 15-May-2013. Table 3 shows the distribution of the labeled UserIDs[5].

Table 3: Number of Labeled and Unlabeled UserIDs as of 15-May-2013

	Trading Group	Number of UserIDs
Labeled	HFT	49
	SB	29
	RET	11
	ST	24
Total Labeled		113
Total Unlabeled		2662
Percent Labeled		4.1%

The labeled set is used to train a supervised learning algorithm to classify User IDs into one of the four categories.

v. Model Training and Evaluation

We evaluate four different types of models for classification described below.

1. Linear Support Vector Machine ("Linear SVM") [5]
2. Linear Support Vector Machine with L1 norm penalty ("Linear SVM with L1 penalty") [6]
3. Radial Basis Support Vector Machine ("Radial Basis SVM") [5]
4. Random Forest [4]

The Support Vector Machine methods (1 to 3 in the list above) are all binary classifiers. These methods are adapted to the multiclass classification problem by considering every possible pairwise model (six binary classifiers in total, corresponding to HFT vs. SB, HFT vs. RET, HFT vs. ST, SB vs. RET, SB vs. ST, RET vs. ST). The final prediction is based on a majority vote. This procedure is known as the one-vs-one approach. See [7] and [8] for discussion on the merits of using this approach.

To evaluate the performance of each type of model we use the following methodology. We first randomly split the labeled dataset into a training and test set. The training set is a stratified sample of 80% of the data; the test set is a stratified sample of the remaining 20% of data. We use 5-fold cross-validation on the training set to tune any model parameters. This procedure avoids selecting a parameter that over-fits the training data. The tuned model is then evaluated on the held out 20% test

[5] Labels were assigned based on prior knowledge of the entity in the course of regulatory and analytic work as well as research into the entity using public data sources.

set. This procedure is repeated 20 times to obtain mean and standard error estimates for out-of-sample performance. This methodology provides an accurate assessment of out-of-sample performance for each type of model. Table 11 (Appendix A) lists the tuning parameters considered for each type of model.

Table 4 reports the mean accuracy obtained by each model on the out-of-sample data. Table 5 through Table 8 shows the mean confusion matrix for each model type. The results show that all model types successfully learn an accurate classification rule. We choose to use a Linear SVM since it obtains a high degree of accuracy with a simple linear hypothesis space.

For the final model, using a Linear SVM, we use separate penalty parameters for each pairwise classification problem. These parameters are selected using 8-fold cross-validation on 100% of the data. The final models are retrained on 100% of the data using the optimal parameters identified.

Table 4: Mean Test Accuracy and Standard Error of Each Model Type on 20 Runs of Out-of-Sample Data

Model Type	**Out-of-sample accuracy on labeled examples**	**Standard Error**
Linear SVM	99.6	0.3
Linear SVM with L1 penalty	99.6	0.3
Radial Basis SVM	98.5	0.6
Random Forest	98.5	0.6

Table 5: Average Confusion Matrix for Linear SVM Model (in Percent)

		Predicted			
		HFT	**SB**	**RET**	**ST**
True	**HFT**	100	0	0	0
	SB	0	98.3	1.7	0
	RET	0	0	100	0
	ST	0	0	0	100

Table 6: Average Confusion Matrix for Linear SVM with L1 Penalty Model (in Percent)

		Predicted			
		HFT	**SB**	**RET**	**ST**
True	**HFT**	100	0	0	0
	SB	0	98.3	1.7	0
	RET	0	0	100	0
	ST	0	0	0	100

Table 7: Average Confusion Matrix for Radial Basis SVM Model (in Percent)

		Predicted			
		HFT	**SB**	**RET**	**ST**
True	**HFT**	100	0	0	0
	SB	4.2	95.0	0.8	0
	RET	0	2.5	97.5	0
	ST	0	0	0	100

Table 8: Average Confusion Matrix for Random Forest Model (in Percent)

		Predicted			
		HFT	**SB**	**RET**	**ST**
True	**HFT**	100	0	0	0
	SB	5.8	94.2	0	0
	RET	0	0	100	0
	ST	0	0	0	100

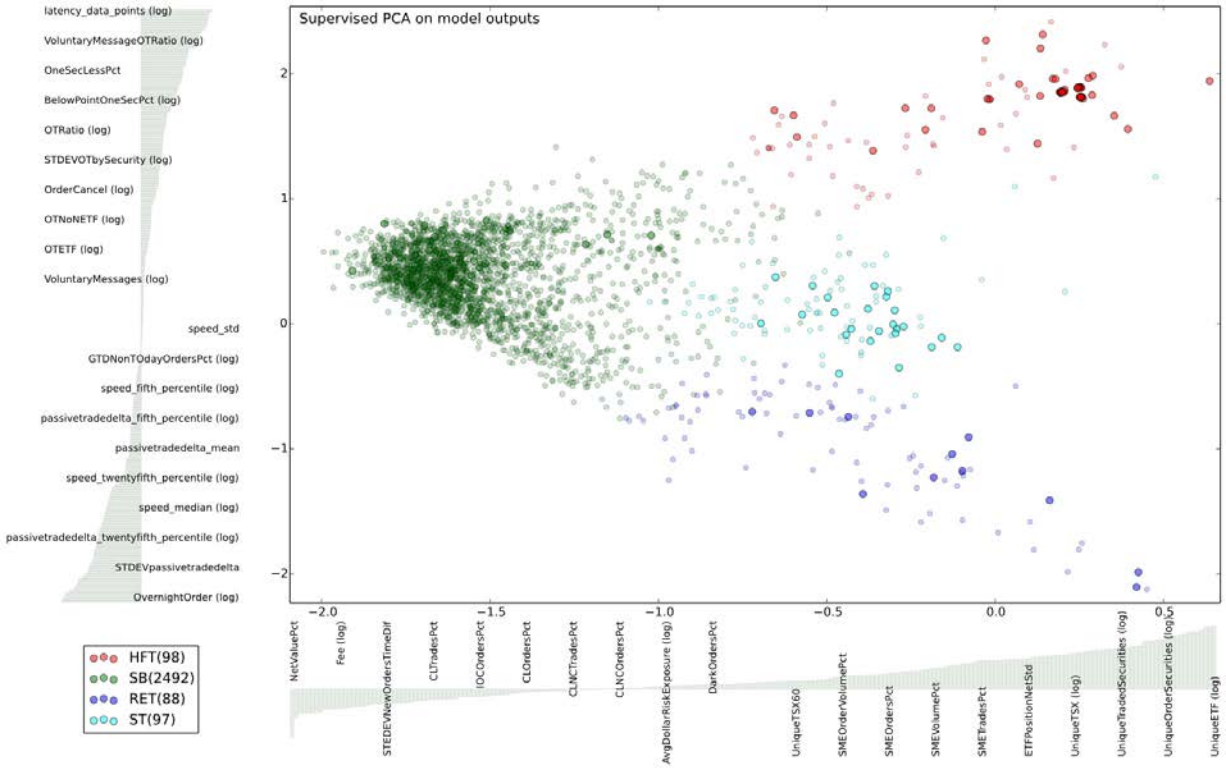
vi. Classification of UserIDs

We use the final trained Linear SVM model to predict labels for all User IDs for each month in the study period. Table 9 shows the number of User IDs identified for each trading group on 15-May-2013. The relative size of each group is consistent with our labeled set and stable month over month. Figure 3 projects the classified UserIDs onto a two-dimensional space using the Supervised PCA algorithm for dimensionality reduction and shows that the output of the classifier produces well separated groups of UserIDs.

Table 9: Distribution of Predicted UserIDs as of 15-May-2013

Trading Group	**Number of UserIDs**
HFT	98
SB	2492
RET	88
ST	97
Total	2775

Figure 3: Visualization of UserIDs on 15-May-2013 Using the Supervised PCA Method [9]



Supervised PCA is used primarily for visualization purposes and projects the classes predicted by the linear SVM onto a two dimensional space. The weighting of the input dimensions with respect to each component are visualized with green bars.

vii. Stability of Classification Over Time

For the purposes of this evaluation and for calculating metrics in Appendices B through F, predictions were made for each month of our review period based on the averaged feature set representing the month's activity. This feature smoothing is necessary to smooth out short-term variability in classifications. However we do expect some variability in classification given that UserIDs phase in and out of existence and because some UserIDs could be representing multiple clients or strategies. The variability here refers to a UserID's predicted class changing over the study period.

A total of 3436 unique UserIDs were segmented over the four month period. The majority of UserIDs (98.2%) were in the same segment each month. The remaining 62 UserIDs (1.8%) were assigned to more than one segment over the four month period. The impact of these ambiguous UserIDs ranges from 2% to 2.5% of monthly volume. Given the overall stability of our classification both in terms of populations of UserIDs as well as the volume represented by those UserIDs, we allowed a UserID to keep the predicted label for each month.

3. Characterizing User Segments

In this section we present some aggregate statistics and interactions amongst the population of all UserIDs classified by the above process. Appendices B through E outline these statistics in detail. Figure 4 shows key trading statistics for all UserID segments. As expected we see the HFT group is the primary contributor to number of order messages with an overall contribution of 91%. On a volume basis, the HFT segment contributes about 17%. Figure 6 shows the percentage of passive volume by each user segment and shows the HFT group to be predominately passive (70% passive). This is in line with our expectations based on Canadian market structure. Appendices C and E show the interactions amongst the groups. HFT, which is the counterparty to 29% of all trading volume, tends to be passive regardless of their trading counterparty.

We further quantify each group's trading activity in terms of transaction costs. We proxy this via volume-weighted effective spreads and realized spreads on all TSX60 securities. Details of how these measures are calculated are included in Appendix F. A more elaborate discussion of these measures can also be found in [10, 11]. The effective spread is calculated for each trade and captures the spread paid (active) or received (passive) relative to the midpoint price of the security. The realized spread captures any adverse selection cost (price impact) to a trader by considering the midpoint price 5 minutes after the trade[6]. Figure 8 shows mean values for these measures for each group over the study period.

Corresponding to their overall passive order flow, we see HFT earning an effective spread of 1.71 bps. The retail group, on the other hand, pays an effective spread of 1.11 bps whereas the SB group pays a smaller effective spread of 1.03 bps. Looking at realized spreads, which can be viewed as net profits for liquidity providers [11], we see the HFT group earned a realized spread of 0.14 bps (less than the effective spread) suggesting that this group is mostly non-directional. On average, HFT earns a further 0.77 bps via net rebates or approximately 5.5 times the potential revenue from simple liquidity provisioning. In the SB group, where we expect more informed flows to exist, the price impact is positive, compared to all other groups which have negative price impacts. The SB price impact is larger than their effective spread costs, and results in earned realized spreads of 0.46 bps. Again these measures in aggregate are in line with our expectations.

It should be noted that these measures serve to characterize the groups in aggregate and to show our classification methodology is producing results that are consistent with our expectations. However, in this study, we do not attempt to characterize the impact of these groups on market quality, integrity or efficiency.

4. Summary and Future Steps

[6] This is consistent with realized spread disclosure under the Securities and Exchange Commission (SEC) rule 605 and the academic literature [12, 10].

We have presented a method of segmenting market participants at a UserID level. The supervised learning approach to segmentation that we have outlined takes advantage of both the rich dataset available to IIROC and our knowledge of market participants. The summary statistics describing the classified groups are in line with our expectations and provide further validation of our methodology for identifying different groups. This method improves upon our previous approaches and represents a solid step towards our long term objective. We intend to continue building upon this work in a number of ways.

First, we intend to apply our model over an extended time horizon and operationalize its use for further studies on each group's impact on key market quality, efficiency and integrity measures.

Second, we will continue to look for means of extending our approach to identify sub-groups within each segment. As others have noted, some groups such as HFT are heterogeneous. To gain insight at a finer granularity we will explore methods by which we can incorporate features that capture intra-day market dynamics and behaviours.

Finally, we will continue to refine the methodology outlined in this paper. For instance, we note that for this work, we have used a supervised learning algorithm where the training set was labeled using domain expertise. We would like to explore more collaborative ways of labeling this dataset using multiple sources of domain expertise. Further, while we saw classification rates improve when we used a large feature set over a small feature set, we see value in reducing the number of features (attributes) for the purposes of classification.

Our long term goal is to go beyond the multiclass classification of users described in this paper, and to be able to identify the trading patterns or strategies that differentiate one group from another and understand the impact of each group on the markets.

5. Notes to the Appendices

Table 10 provides some information concerning the analysis conducted in each of the following Appendices to assist interpretation.

Table 10: Notes to the Appendices

Notes:	**Appendix**				
	B	**C**	**D**	**E**	**G**
UserID segments are assigned based on a single prediction per month for each UserID; the prediction is based on the averaged feature set representing the month's activity	x	x	x	x	x
Statistics are aggregated by day for each segment, and then a daily average is computed; average volume and average value refer to traded volume and traded value	x	x	x	x	
Trading activity on all marketplaces in all listed securities between 00:00 and 24:00 is included	x	x	x	x	
The AAPP category describes trading which is either active-active or passive-passive; examples of this type of trading include crosses, MOC trading, opening trades and trading on MatchNow			x	x	
Volume weighted averages were calculated for each UserID segment daily, and then a daily average is computed					x
Analysis restricted to trading activity in TSX60 securities only (on all traded marketplaces) between 09:30 and 16:00					x
Only trades which have both an active and passive side were included; AAPP trades (see above) were excluded					x
Trade volume, value and numbers are double counted, in that the buyer and seller each count the trade	x		x		x
Trade volume, value and numbers are single counted; each trade is counted only once		x		x	

6. Appendix A

Table 11 lists the tuning parameters considered for each type of model.

Table 11: Tuning Parameters for Each Model Type

Model Type	**Parameter**	**Set of Value(s)**
Linear SVM	Penalty of misclassification (C)	2^[-15,-14,…,14,15]
Linear SVM with L1 penalty	Penalty of misclassification (C)	2^[-15,-14,…,14,15]
Radial Basis SVM	Penalty of misclassification (C)	2^[-15,-14,…,14,15]
	Radial Basis Kernel Width (σ^2)	2^[-10,-9,…,9,10]
Random Forest	Number of Trees	500

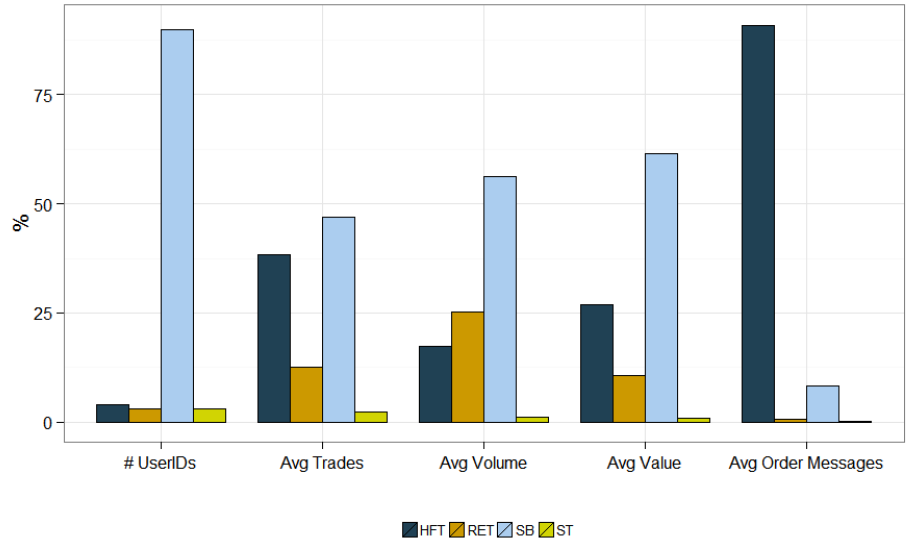
7. Appendix B: Summary Statistics by UserID Segment

Table 12 and Figure 4 report summary statistics by UserID segment.

Table 12: Daily Average Summary Statistics – Percentage by UserID Segment

UserID Segment	Number of UserIDs	Average Volume	Average Value	Average Trades	Average Number of Orders	Average Order to Trade Ratio
HFT	4%	17%	27%	38%	91%	55.4
RET	3%	25%	11%	13%	1%	1.1
SB	90%	56%	61%	47%	8%	4.1
ST	3%	1%	1%	2%	0%	3.2

Figure 4: Daily Average Summary Statistics – Percentage by UserID Segment



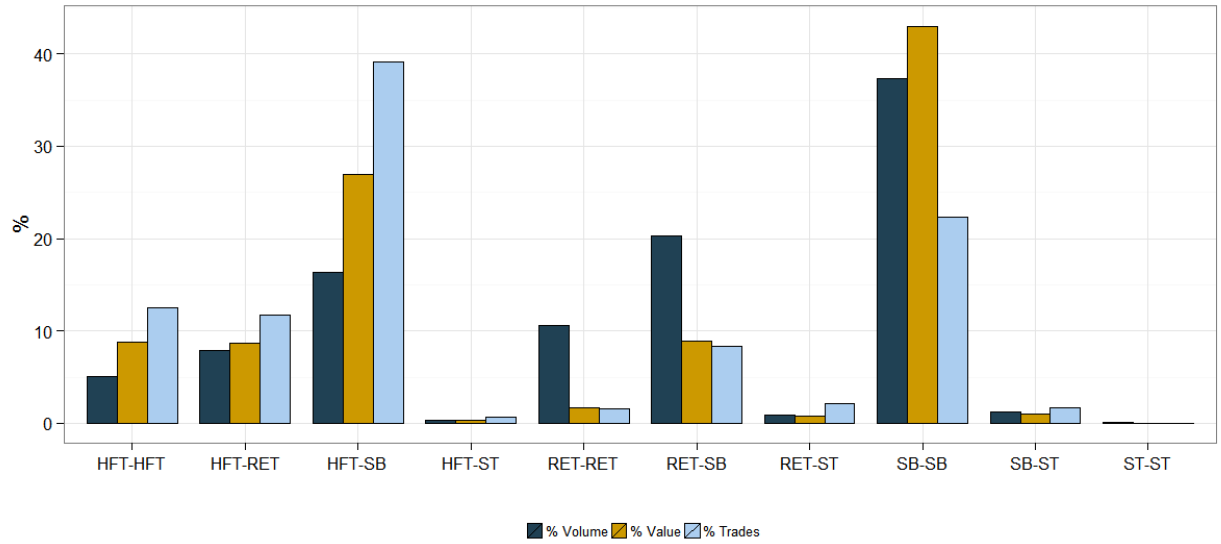
8. Appendix C: Trading Statistics by UserID Segment and Counterparty

Table 13 and Figure 5 report summary trading statistics by UserID segment and counterparty.

Table 13: Daily Average Volume, Value and Number of Trades – Percentage by UserID Segment and Counterparty

UserID	Counter UserID	Average Volume	Average Value	Average Trades
HFT	HFT	5%	9%	12%
HFT	RET	8%	9%	12%
HFT	SB	16%	27%	39%
HFT	ST	0%	0%	1%
RET	RET	11%	2%	2%
RET	SB	20%	9%	8%
RET	ST	1%	1%	2%
SB	SB	37%	43%	22%
SB	ST	1%	1%	2%
ST	ST	0%	0%	0%

Figure 5: Daily Average Volume, Value and Number of Trades – Percentage by UserID and Counter UserID Segment



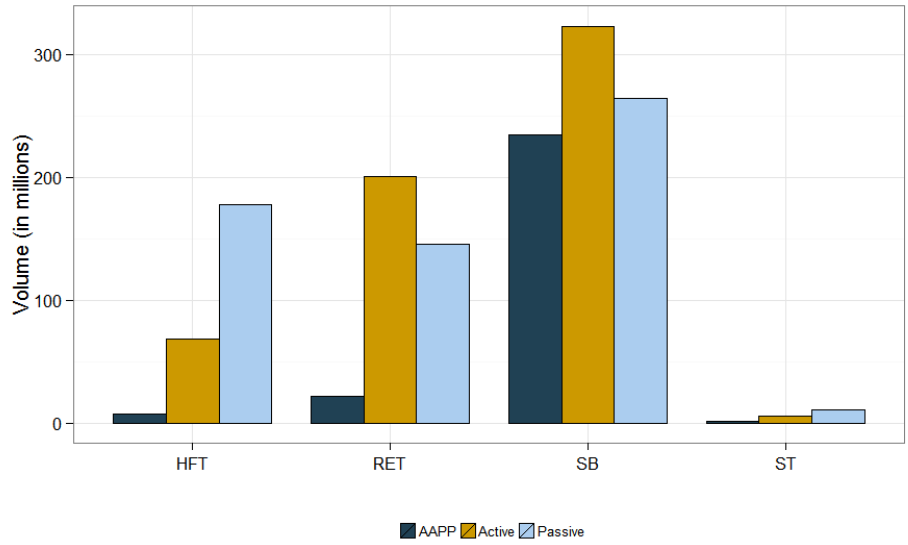
9. Appendix D: Active vs. Passive Volume by UserID Segment

Table 14 and Figure 6 report the amount and proportion of active and passive trading by UserID segment.

Table 14: Daily Average Percentage Active / Passive Volume – by UserID Segment

UserID Segment	% AAPP	% Active	% Passive
HFT	3%	27%	70%
RET	6%	54%	40%
SB	29%[7]	39%	32%
ST	9%	30%	61%

Figure 6: Daily Average Absolute Active / Passive Volume – by UserID Segment



[7] The large AAPP percentage is due to intentional crosses, which is expected for this group.

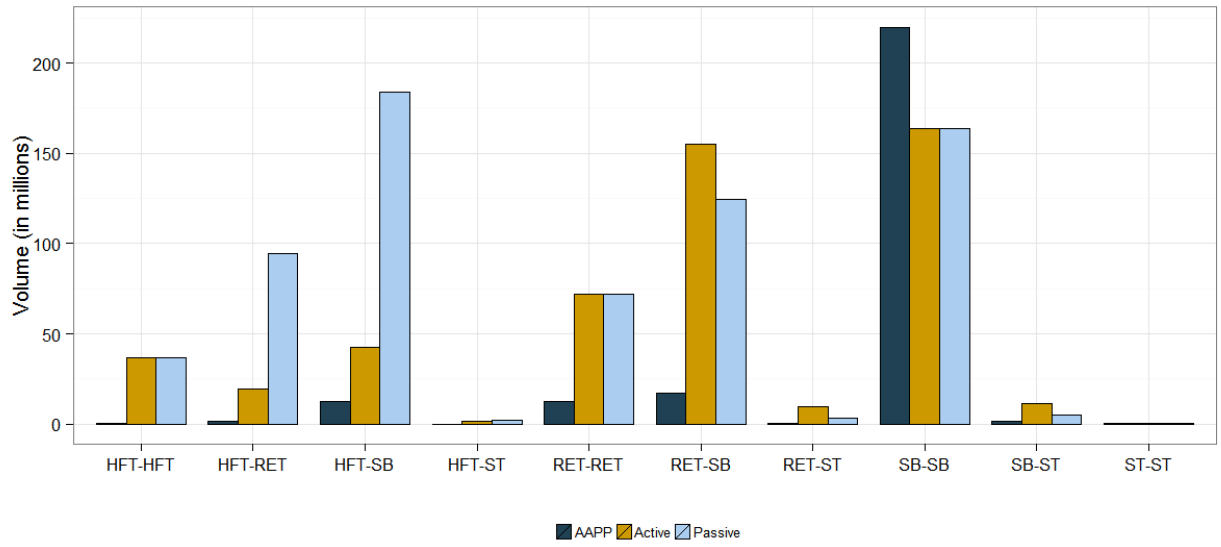
10. Appendix E: Active vs. Passive Volume by UserID Segment and Counterparty

Table 15 and Figure 7 report the amount and proportion of active and passive trading by UserID segment.

Table 15: Daily Average Percentage Active / Passive Volume – by UserID Segment and Counterparty

UserID	Counter UserID	% AAPP by UserID	% Active by UserID	% Passive by UserID
HFT	HFT	0%	50%	50%
HFT	RET	1%	17%	82%
HFT	SB	5%	18%	77%
HFT	ST	2%	37%	61%
RET	RET	8%	46%	46%
RET	SB	6%	52%	42%
RET	ST	5%	69%	26%
SB	SB	40%	30%	30%
SB	ST	8%	64%	28%
ST	ST	36%	32%	32%

Figure 7: Daily Average Absolute Active / Passive Volume – by UserID Segment and Counterparty



11. Appendix F: Trading Costs by UserID Segment - Definitions

Effective Half Spread ("ES")

ES measures the difference between the trade price (P_{it}) and the current value of the security, proxied by the mid-point of the spread at the time of the trade (V_{it}) and is scaled by the midpoint at the time of the trade. The formula is as follows, where D_{it} is an indicator variable which takes the value +1 for the buyer and -1 for the seller [10]:

$$ES_{it} = D_{it} * \frac{(P_{it} - V_{it})}{V_{it}}$$

From the point of view of the buyer, the effective spread is positive if the trade price is higher than the midpoint (for example, when an active buy order crosses the spread), and negative if the trade price is lower than the midpoint (for example, when an active sell order crosses the spread). From the point of view of the seller, the effective spread is positive if the trade price is lower than the midpoint, and negative if the trade price is higher than the midpoint.

Price Impact ("PI")

PI measures the difference between the future value of the security, proxied by the mid-point of the spread 5 minutes after the trade (V_{it+5}) and the current value of the security, proxied by the mid-point of the spread at the time of the trade (V_{it}) and is scaled by the current value of the security. The formula is as follows:

$$PI_{it} = D_{it} * \frac{(V_{it+5} - V_{it})}{V_{it}}$$

From the point of view of the buyer, the price impact is positive if the price goes up after a trade, and negative if the price goes down. From the point of view of the seller, the price impact is positive if the price goes down after a trade, and negative if the price goes up.

Realized Half Spread ("RS")

RS measures the difference between the trade price (P_{it}) and the future value of the security, proxied by the midpoint of the spread 5 minutes after the trade (V_{it+5}), and is scaled by the current value of the security. The formula is as follows:

$$RS_{it} = D_{it} * \frac{(P_{it} - V_{it+5})}{V_{it}}$$

From the point of view of the buyer, the realized spread is positive if the future midpoint price is lower than the trade price, and negative if the future midpoint price is higher than the trade price. From the point of view of the seller, the realized spread is positive if the future midpoint price is higher than the trade price, and negative if the future midpoint price is lower than the trade price.

The three measures are related as follows:

$$RS = ES - PI$$

Marketplace Costs (Fees and Rebates) ("MC")

MC scales the fee paid or rebate earned by the current value of the security so that it can be compared to (or incorporated into) the realized and effective spreads. The formula is as follows, where R_{it} takes a value from Table 16 below, based on the marketplace where the trade took place, and whether the UserID was on the active or passive side of the trade:

$$MC_{it} = \frac{-1 * R_{it}}{V_{it}}$$

Table 16 is based on a simplification of publicly available information concerning historical and current fee structures. This simplification is sufficient because the analysis has been restricted to trades in the TSX60 which had an active and passive side. In the table below, rebates are positive numbers and fees are negative numbers (as published by the marketplaces). Effective spread and realized spread are calculated such that benefits are negative numbers and costs are positive numbers. This is the reason that Marketplace Costs transform rebates and fees by -1. Negative marketplace costs indicate that the trader is receiving net rebates; positive marketplace costs indicate that the trader is paying net fees.

Table 16: Estimated Rebate and Fee Structure for the Period (in Dollars)

Marketplace	Active	Passive
ALF	-0.0028	0.0025
CHX	-0.0029	0.0025
CNQ	-0.0025	0.002
CX2	0.001	-0.0014
ICX	-0.0015	-0.0015
LIQ	-0.01	-0.01
OMG	-0.0006	0
PTX	-0.0025	0.002
TCM	-0.001	-0.001
TMS	-0.0009	0.0005
TSX	-0.0035	0.0031
TSXV	-0.0035	0.0031

12. Appendix G: Trading Costs by UserID Segment - Findings

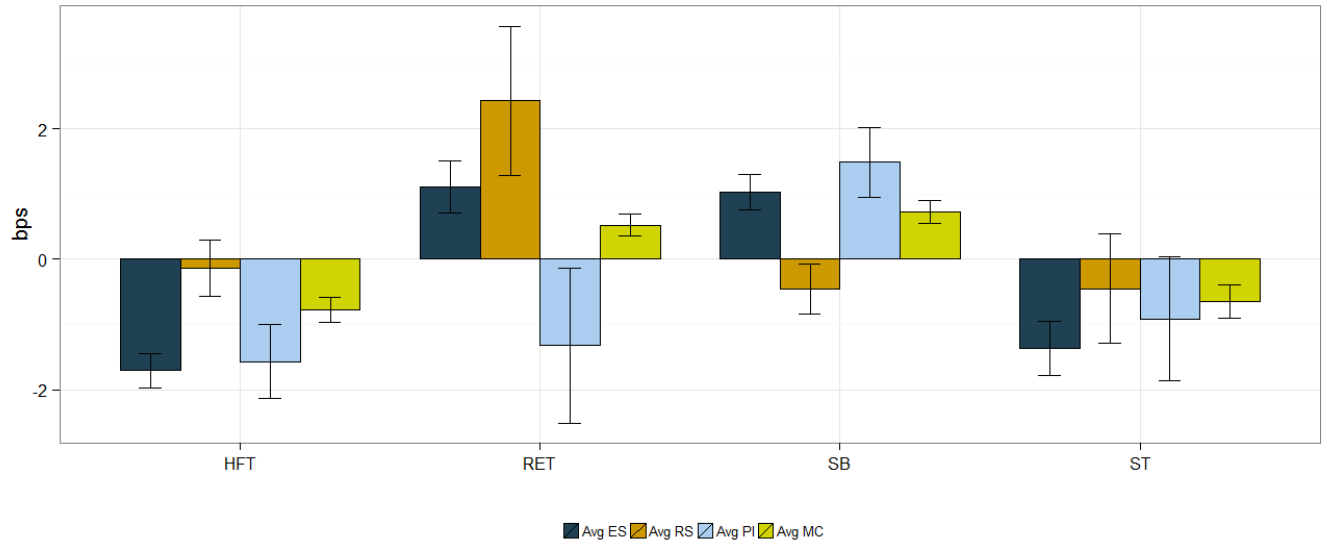
Table 17 and Figure 8 report trading cost measures attributed to UserID segment.

ES, RS, PI and MC were calculated for the buyer and seller of each trade, and attributed to the relevant UserID segment. Volume weighted averages were calculated for each UserID segment daily. Table 17 and Figure 8 show the average spreads, price impact and marketplace costs. Error bars show one standard deviation from the mean.

Table 17: Cost Measures (in bps) by UserID Segment

UserID Segment	Average Effective Half Spread	Average Realized Half Spread	Average Price Impact	Average Marketplace Cost
HFT	-1.71	-0.14	-1.57	-0.77
RET	1.11	2.43	-1.32	0.53
SB	1.03	-0.46	1.49	0.72
ST	-1.36	-0.45	-0.91	-0.65

Figure 8: Cost Measures by UserID Segment



13. References

[1] IIROC, "The HOT study." Discussion paper, 2012.

[2] IIROC, "Market quality in a rapidly changing environment," Investment Industry Regulatory Organization of Canada, 2013. OSC-IIROC Market Structure Conference - The Canadian Equity Market: Structural Challenges Amidst Rapid Change.

[3] V. N. Vapnik, *Statistical learning theory*. Wiley, 1 ed., 1998.

[4] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[6] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines.," in *ICML*, pp. 82–90, Morgan Kaufmann, 1998.

[7] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Trans. Neur. Netw.*, vol. 13, no. 2, pp. 415–425, 2002.

[8] K. bo Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," in *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*, pp. 278–285, 2005.

[9] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Zolghadri Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recogn.*, vol. 44, no. 7, pp. 1357–1371, 2011.

[10] H. Bessembinder and K. Venkataraman, *Bid-Ask Spreads*. John Wiley & Sons, Ltd, 2010.

[11] K. Malinova, A. Park, and R. Riordan, "Do retail traders suffer from high frequency traders?." Available at SSRN: http://ssrn.com/abstract=2183806 or http://dx.doi.org/10.2139/ssrn.2183806, November 2013.

[12] SEC, *NMS Security Designation and Definitions*. 17 CFR Ch. II 242.600, 2013.