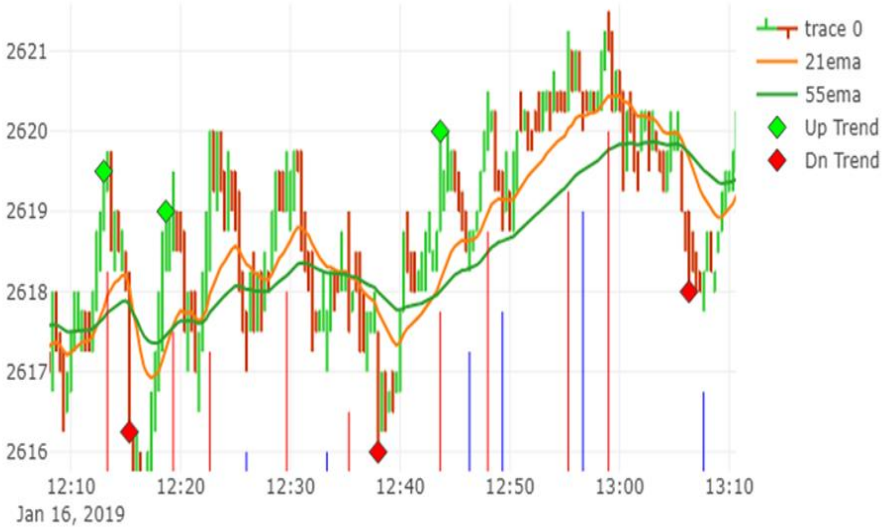
SPRINGBOARD - CAPSTONE 1 - PREDICTING INTRADAY STOCK INDEX FUTURES SWINGS

Tom Preston

AUGUST 7, 2019

AGENDA

- Problem Statement
- The Data
- Exploration
- Random Forest Model
- Results
- Insights



PROBLEM STATEMENT

- SP500 e-Mini volatility has increased over the past few years
- Day trading the e-Mini has increased in popularity
- Several day trading approaches recommend trading short term (1 min) charts for 3 ticks ($37.50 @$12.50 per tick profit) while risking 6 ticks ($75.00) per contract on each trade
- Can we use data science and machine learning to improve on this day trading approach ?

DAY TRADING BASIC CALCULATIONS

- Scenarios A – F compare profit / loss assuming 3 ticks ($37.50) profit & 6 ticks ($75.00) loss over 100 potential trades
- Scenario A – 67% profitability is breakeven approach (before commissions)
- Scenario B - C: 8 more wins or losses greatly impacts profitability
- Scenario D – F: Averaging .5 points more profit per trades improves all scenarios

Scenario A	Trades Per Week	Ticks	P/L
Wins	67	**3**	201
Losses	33	-6	-198
		Net Ticks	3
Scenario B			
Wins	75	**3**	225
Losses	25	-6	-150
		Net Ticks	75
Scenario C			
Wins	60	**3**	180
Losses	40	-6	-240
		Net Ticks	-60

Scenario D	Trades Per Week	Ticks	P/L
Wins	67	**3.5**	234.5
Losses	33	-6	-198
		Net Ticks	36.5
Scenario E			
Wins	75	**3.5**	262.5
Losses	25	-6	-150
		Net Ticks	112.5
Scenario F			
Wins	60	**3.5**	210
Losses	40	-6	-240
		Net Ticks	-30

THE DATA

SP500 E-MINI TRADING DATA FOR JANUARY 2019

- Tick data (1M + ticks)
- Data converted using pandas resampling int Open, High, Low and Close data
- 20 second bar chart (mimics the 687 ticks bar chart for day trading)

ADDITIONAL DATA:

- 21 bar EMA (exponential moving average) – pd.Series calculation
- 55 bar EMA (exponential moving average) – pd.Series calculation

CUSTOM DEVELOPED CALCULATIONS:

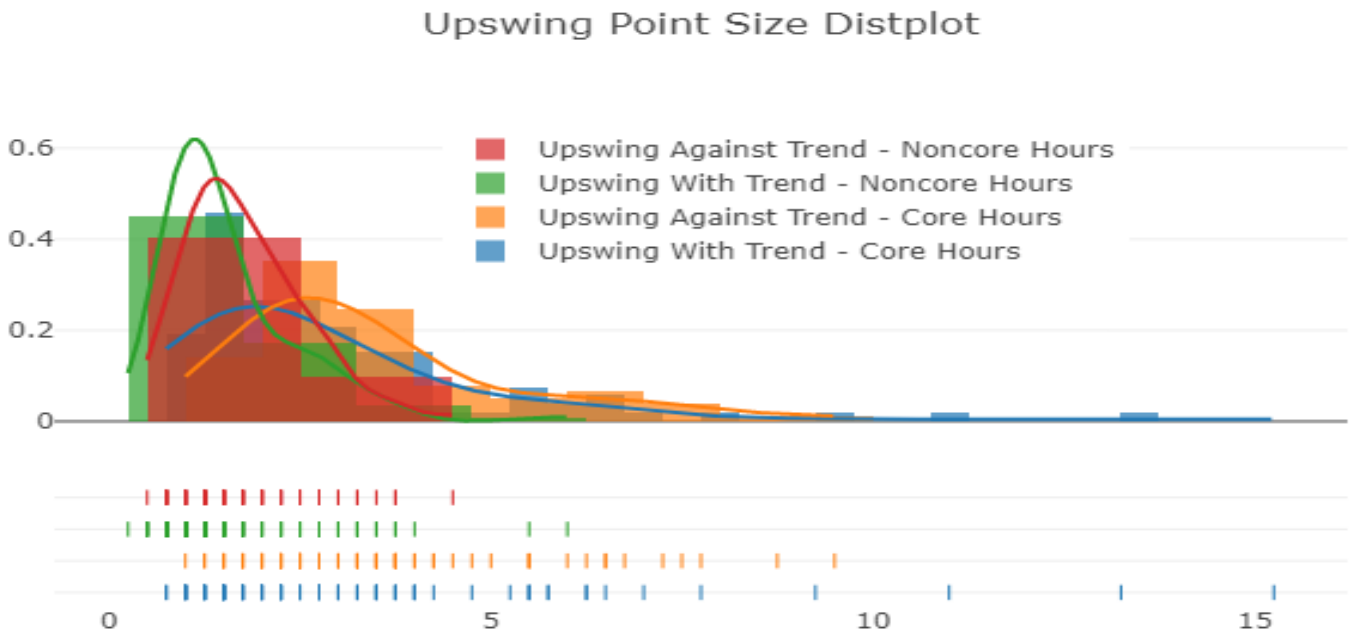
- Swing (up or down)
  - The size of the swing (in points) from the beginning to the end of the swing
  - The number of total bars in a swing
- Trend (up or down) – Trend is up when price closes 1.5 point above 55 EMA and vice versa
- Trend Move (up or down)
  - The MFE (most favorable excursion) – maximum open profit for a trend move
  - The MAE (most adverse excursion) – maximum drawdown for a trend move
  - The number of bars to hit the MFE and MAE

E-MINI PRICE CHART EXAMPLE

E-MINI EDA

EDA FOR PRICE SWINGS AND TREND MOVES

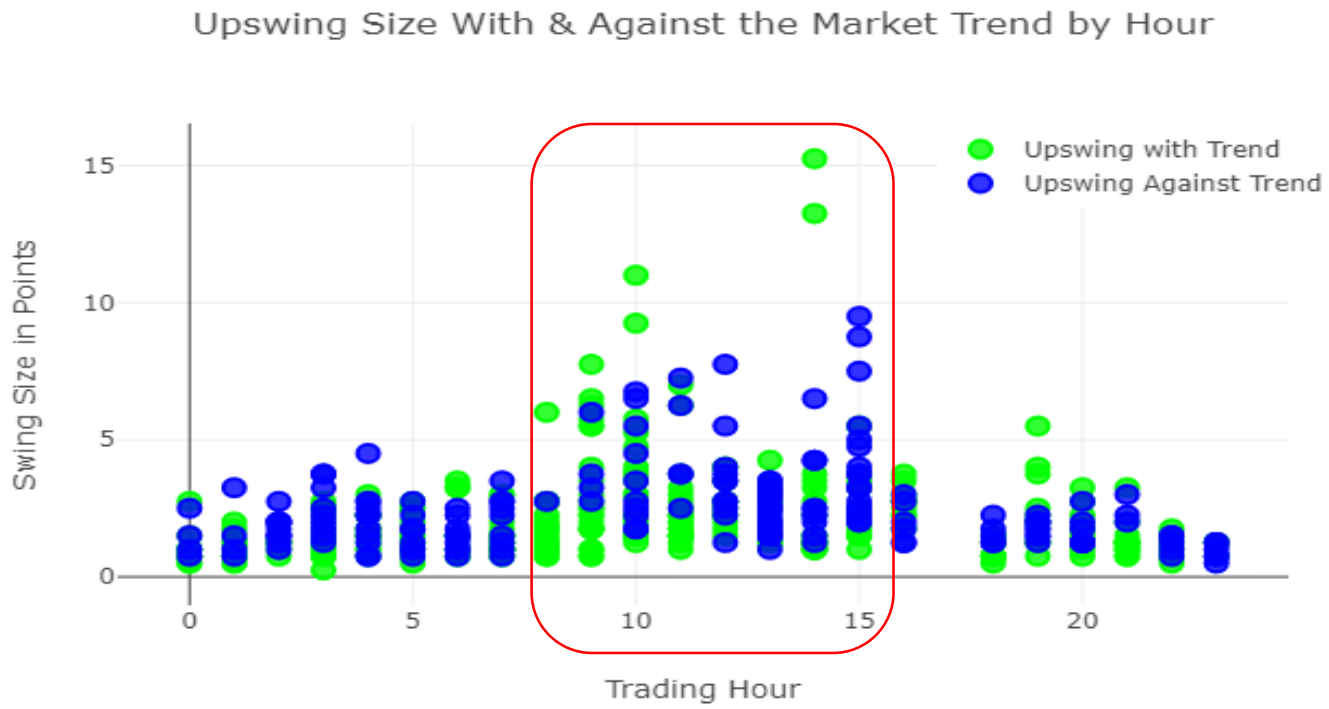
- Move size in points and move length in number of bars
- Data separated into core hours (9 am – 3 pm EST) and non core hours (4 pm – 8 am EST)
- Swings with and against the trend analyzed



E-MINI CORE TRADING HOURS

EDA FOR PRICE SWINGS AND TREND MOVES

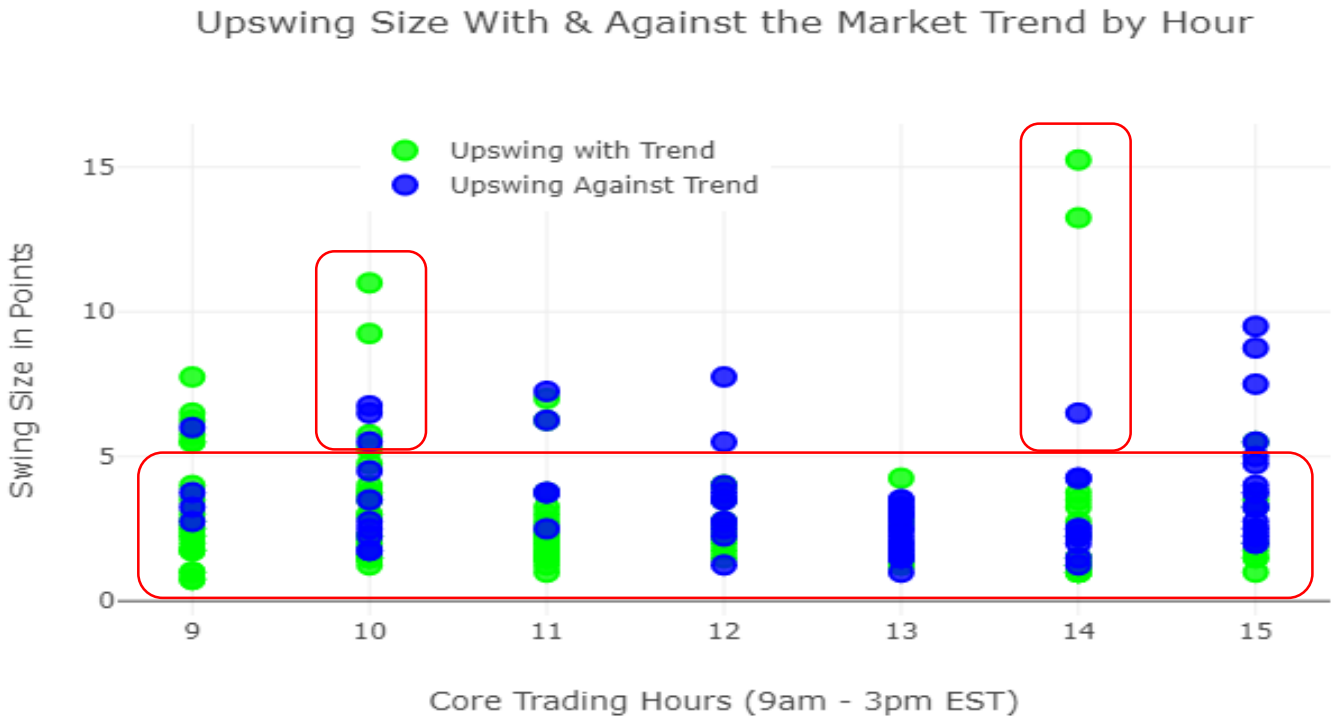
- Core trading hours (9 am – 3 pm EST) shows larger size and longer length moves
- Trading with the trend are typically larger



E-MINI CORE TRADING HOURS CONT.

EDA FOR PRICE SWINGS AND TREND MOVES

- Core trading hours average 5 point moves
- 10 am and 2 pm EST often have larger moves



SUMMARY SWING ANALYSIS

EDA FOR PRICE SWINGS (956 OBSERVATIONS)

- 50% of all swings are 1.75 points or larger therefore .75 pt profit target is reasonable
- Core trading hours 50% leverage for all swings average 2.5 points
- Swing length are similar across all swing types – marker keeps same swing structure

Swing Type	Sample Size	Move Size in Points (% Levels)			# of 20 second bars (% Levels)		
		20%	**50%**	**80%**	**20%**	**50%**	**80%**
All Swings	956	1.00	1.75	2.75	4	9	7
Up swings – Core Hours – with trend	106	1.50	2.25	3.75	4	8	16
Up swings – Non Core Hours – with trend	187	.75	1.25	2.25	4	8	16
Up swings – Core Hours – against trend	77	2.00	3.00	4.50	5	10	20
Up swings – Non Core Hours – against trend	108	1.25	1.5	2.5	5	11	18
Down swings – Core Hours – with trend	55	1.75	2.75	4.00	3	6	12
Down swings – Non Core Hours – with trend	96	1.00	1.50	2.00	4	8	15
Down swings – Core Hours – against trend	128	1.5	2.5	3.75	4	8	15
Down swings – Non Core Hours – against trend	199	1.00	1.25	1.75	5	9	18

SUMMARY TREND MOVE ANALYSIS

EDA for trend moves (93 observations) – Multiple swings make up a trend move

- MFE – trend move potential profit 50% level is 2.0 points
- Length of trend is much larger during non-core hours meaning price is not as volatile

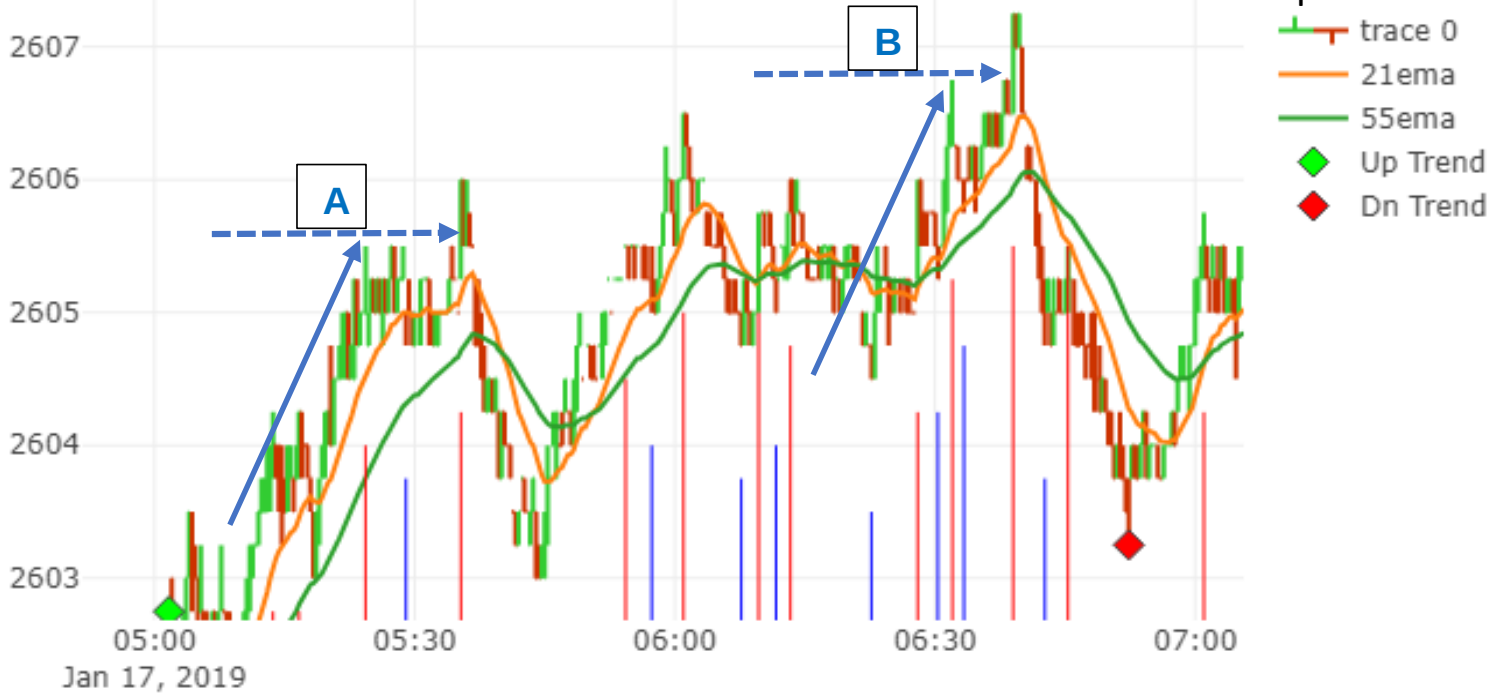
Trend Moves	Sample Size	Most Favorable Excursion Points (% Levels)			20 second bars (% Levels)		
		20%	**50%**	**80%**	**20%**	**50%**	**80%**
All Trend Moves	93	1.0	2.0	5.0	3	27	97
All trends – Core Hours	61	.5	1.75	5.0	2	16	46
All trends – Non Core Hours –	32	.75	1.25	2.25	17	68	225

- MAE – trend move drawdown 50% level is (1.0) points
- Data supports using limit entry approach for entering trend trades (i.e. enter on price pullback

Trend Moves	Sample Size	Most Adverse Excursion Points (% Levels)			20 second bars (% Levels)		
		20%	**50%**	**80%**	**20%**	**50%**	**80%**
All Trend Moves	93	(.50)	(1.0)	(2.0)	3	27	97
All trends – Core Hours	61	(.25)	(1.0)	(1.75)	1	2	15
All trends – Non Core Hours –	32	(.50)	(1.25)	(2.00)	2	15	37

RANDOM FOREST APPLICATION

- Price swing analysis provides reasonable ranges for swing size and length
- Predicting if the price swing will exceed previous swing high / low establishes trading targets
- The goal is to predict if the swings will exceed previous swing extrement (i.e. A and B high points) for all swings
- Random Forest will be used to determine the most valuable features for this prediction



RANDOM FOREST MODEL

Our data is reduced to these columns:

- Breakout (our target variable) – 1 or 0
- 21 EMA & 55 EMA
- Retrace = the percentage retracement of the previous swing
- Swing (-1 = downswing, 1 = upswing)
- Swing_len – number of 20 second bars in swing
- Swing_type (4 types of swings – Long and Short, with and against the trend)
- Trend (1 = up trend, -1 = down trend)
- Hour = trading hour (24 hour scale)
- Hour_type (9 am – 3 pm = 1 (“core” hours) and 4 pm – 8am = 0 (“noncore” hours)
- Sample Data below:

	21EMA	55E	Brkout	retrace	swing	swing_len	swing_type	trend	hour	hour_type
0	2614.33	2614.14	1.0	1.00	-1	2	4	1.0	3	0
1	2614.33	2614.15	1.0	2.00	1	2	1	1.0	3	0
2	2614.28	2614.14	1.0	1.50	-1	4	4	1.0	3	0
3	2613.97	2614.03	0.0	2.33	1	7	1	1.0	3	0
4	2614.03	2614.04	0.0	0.86	-1	4	4	1.0	3	0

RANDOM FOREST RESULTS

- 80 % train and 20% test data
- Accuracy: 0.65, Precision: .66, Recall 0.69

Predicted Result	0.0	1.0
Actual Result		
0.0	53	36
1.0	32	71

Feature Importance (top 3 marked)

- **('21EMA', 0.16784893965323588), <<<<<<<<<<<**
- ('55E', 0.1474478287573294),
- **('retrace', 0.29807399850023714), <<<<<<<<<<<**
- ('swing', 0.015242575697336433),
- **('swing_len', 0.19725555099071168), <<<<<<<<<<<**
- ('swing_type', 0.034379451027185 43),
- ('trend', 0.015427164478980612),
- ('hour', 0.11245784433977743),
- ('hour_type', 0.011866646555205956)]

RESULTS / FINAL INSIGHTS

- Random Forest Model:
  - 66% accuracy is solid start but more features are needed to improve model
  - Top 3 features (21 EMA, retracement, and swing length)
    - Need to add decision tree to understand how each feature is being used
    - Retracement feature assumed to mean smaller pullbacks signal market strength therefore predicting previous swing peak to be exceeded
    - Swing Length should be removed – the length of the swing ies only know after the swing is completed
- SP500 eMimi Final Insights:
  - Core trading hours have more volatility
  - Approximately 10 – 1 ratio of trading swings to trend moves – while trend moves go longer, proper filtering of trade swings provides more opportunities
  - Trading with trend makes sense but intraday swings can be traded with and against the trend
  - Trend swing analysis supports limit entries, additional study needed on price swings
  - Additional analysis around price retracement is next areas to study

LESSONS LEARNED

- Accuracy Scores can be misleading
  - Review all the model statistics (f1 score, recall, confusion matrix) to assess the model
  - Review from the business problem context
    - Predicting more false positives (predict default but consumer pays on time) is better than false negatives (predicting consumer pays on time but actually defaults)

- Technical challenges – memory issues
  - 30,000 row, 25 columns (12 of which are floats)
  - Scalar method
  - Ran 10K with all float columns w/o scalar

- Lessons Learned
  - Need additional data (income, credit scores, etc. to better identify potential default payments
  - Most people have late payments – these ebb and flow
  - Each model has parms to experiment with – need to develop more expertise with them
  - Need ability to test larger data sets

SOURCES

- Kaggle Data: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
- New York Federal Reserve Consumer Debt Study: https://www.newyorkfed.org/medialibrary/interactives/householdcredit/data/pdf/HHDC_2017Q4.pdf
- Kaggle Python Code with Graphics code I utilized: https://www.kaggle.com/mahyar511/payment-default-prediction-neural-network