Week 09/16 Summary

1. We tested the agreement rate between Lee Ready method and assigning buy/sell from the closest quotes method.

* We improved the speed to calculate the assigning buy/sell from the closest quotes method.
* The agreement rate is ~60%.

1. We use AMZN 23th data to test the concurrent relationship between price return and SOI. We got the best result with 40+% adjusted R^2 for Lee-Ready method.

* We find the best results by using loop to find the best combination of the three parameters: time bin, threshold value (trades with size above which are excluded), and bucket size.

First decide the range of three parameters to test (start, end, increment)

* Bucket size: (1000,10000,1000)
* Time bin: (15,150,15)
* Threshold size: (1000,5000,1000)

\*Given density of trades size (please see in appendix plot 1)

* Formula: Bucket log price return ~ SOI

\*SOI adjusted by the bucket volume, thus range is [-1, +1]

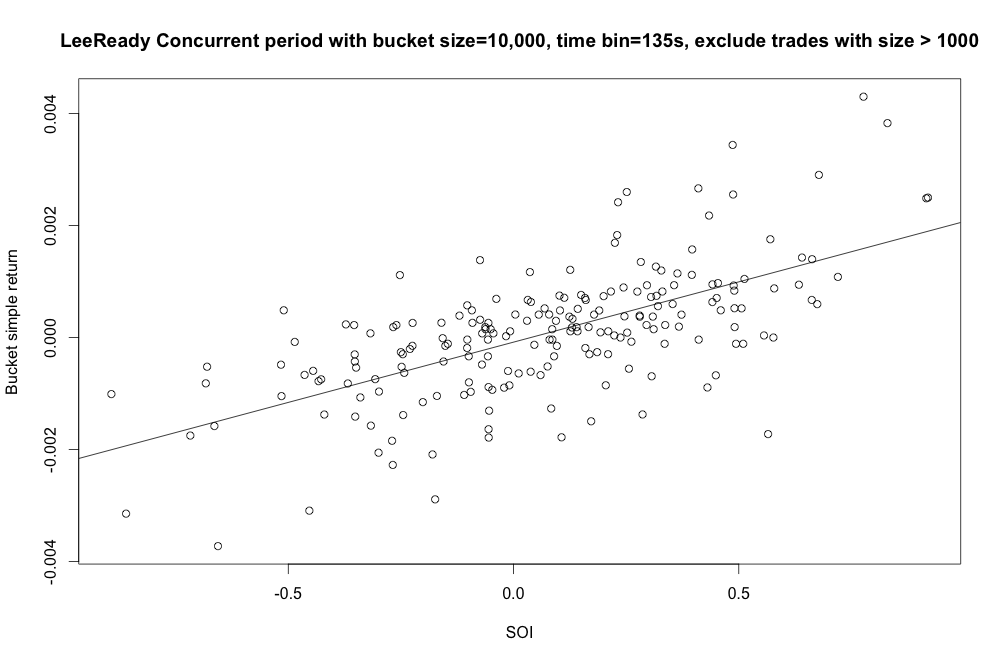
In total, we tested 500 sets of the parameters. 2 of them have R^2 >40%, 54 of them (10.8%) have R^2>30%. Major findings and plots are as the following:

* Best 10 results:

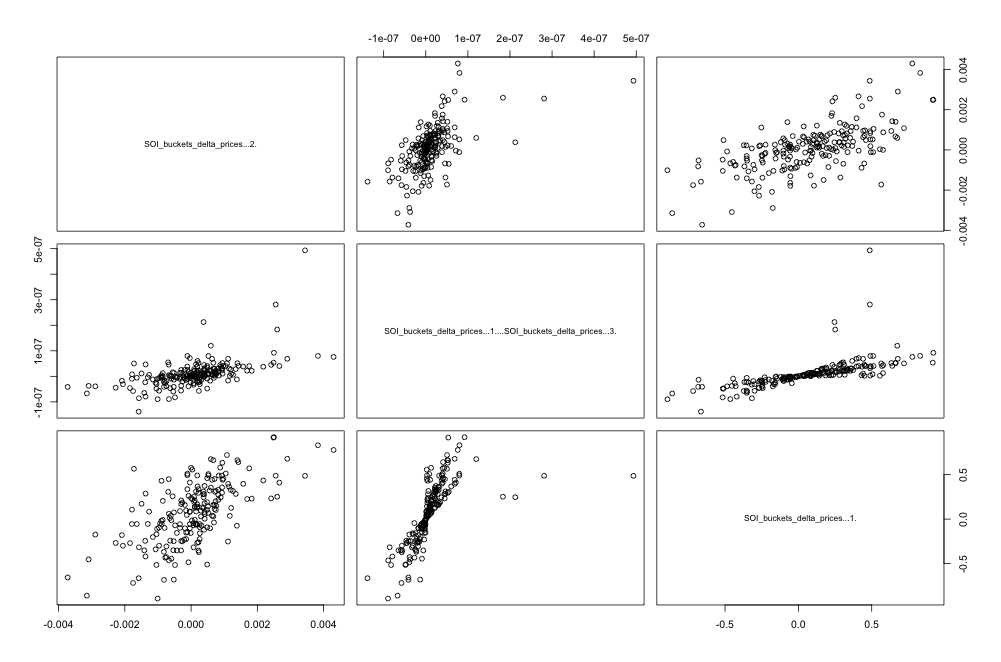
|  |  |
| --- | --- |
| **BucketSize\_TimeBin\_Threshold** | **R^2** |
| 10000 \_ 135 \_ 1000 | 41.30% |
| 10000 \_ 150 \_ 1000 | 41.30% |
| 10000 \_ 120 \_ 1000 | 37.40% |
| 8000 \_ 135 \_ 1000 | 37.10% |
| 9000 \_ 120 \_ 1000 | 36.60% |
| 10000 \_ 105 \_ 4000 | 36.50% |
| 8000 \_ 150 \_ 1000 | 36.20% |
| 10000 \_ 105 \_ 1000 | 36.20% |
| 8000 \_ 120 \_ 1000 | 36.10% |
| 10000 \_ 90 \_ 1000 | 35.60% |

Looking at the best combinations, we find that bucket size over 8000/ time bin over 105/threshold of 1000 generally produce good results. Moreover, it is very interesting that the combination of Bucket Size of 10000 and Threshold 1000 consistently outperforms other combinations. And in this combination, the time bin length is less relevant.

* Take a look at the best result:



Add the cross product of bucket price volatility and SOI as additional independent variable. Variables, as in the sequence below, are: Bucket return, cross product, and SOI.



The plots indicate very good linear relationship between variables.

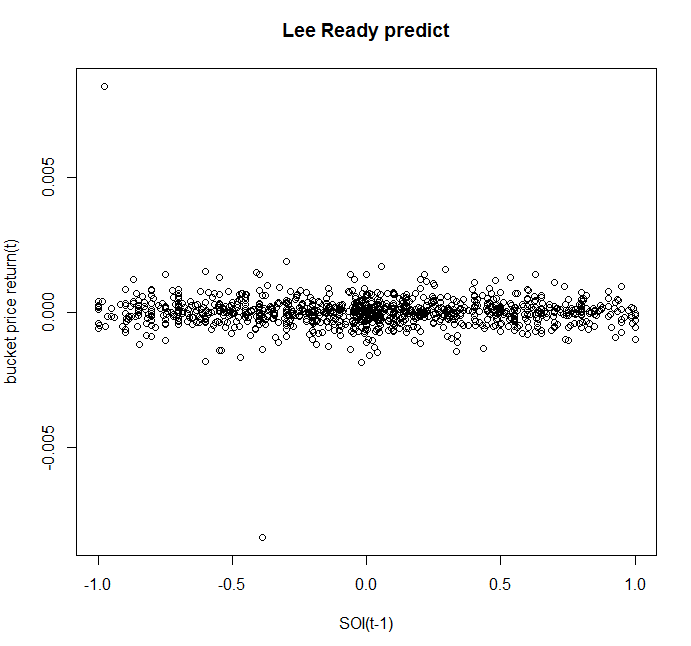
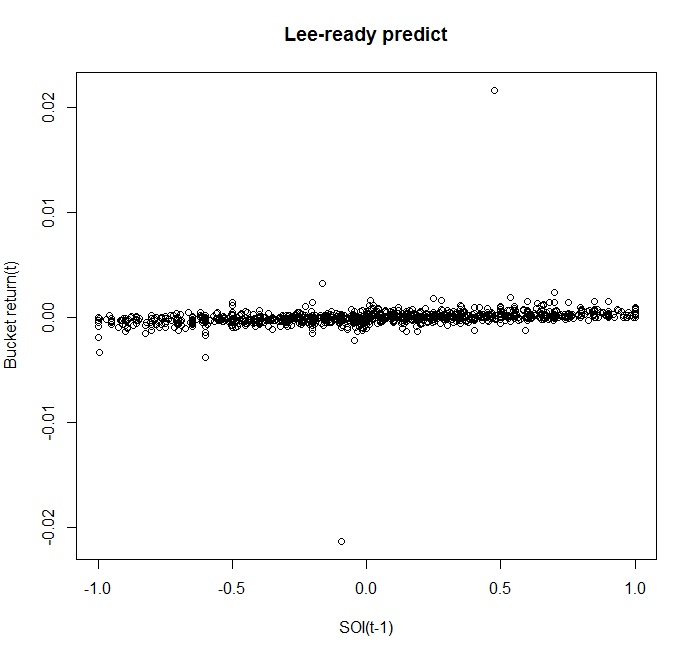
1. Forecast with Lee-ready:

Formula: Bucket returnt ~ SOIt-1

|  |  |
| --- | --- |
| BucketSize\_TimeBin\_Threshold | R^2 |
| 2000 \_ 30 \_ 4000 | **43.6%** |
| 2000 \_ 60 \_ 4000 | **43.6%** |
| 2000 \_ 150 \_ 4000 | **43.2%** |
| 2000 \_ 120 \_ 4000 | **43.1%** |
| 2000 \_ 90 \_ 4000 | **43.1%** |
| 2000 \_ 90 \_ 2000 | **15.4%** |
| 2000 \_ 120 \_ 2000 | **13.8%** |
| 2000 \_ 30 \_ 2000 | **13.0%** |
| 3000 \_ 90 \_ 3000 | **7.8%** |
| 1000 \_ 30 \_ 2000 | **7.7%** |

Though we get some seemingly good result, they are actually inflated by the influential points as the plot below. Thus the result is much weaker than the concurrent one, which is can be intuitively explained, that the order imbalance have direct greater influence on the current bucket.

(2000 \_ 30 \_ 4000) (2000 \_ 90 \_ 2000)

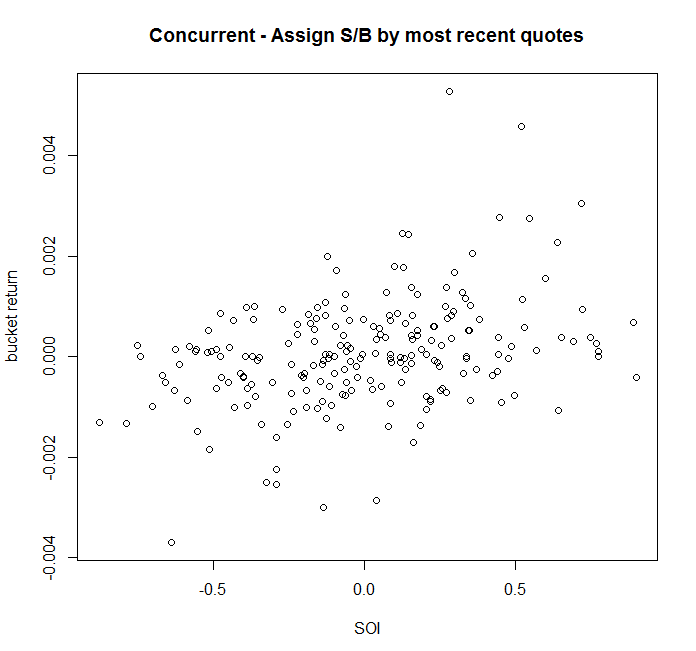


Exclude the influential points, the relationship between the SOI(t-1) and Price return(t) is randomly distributed around zero. This indicates that the predict power of Lee-Ready method calculated SOI is weak.

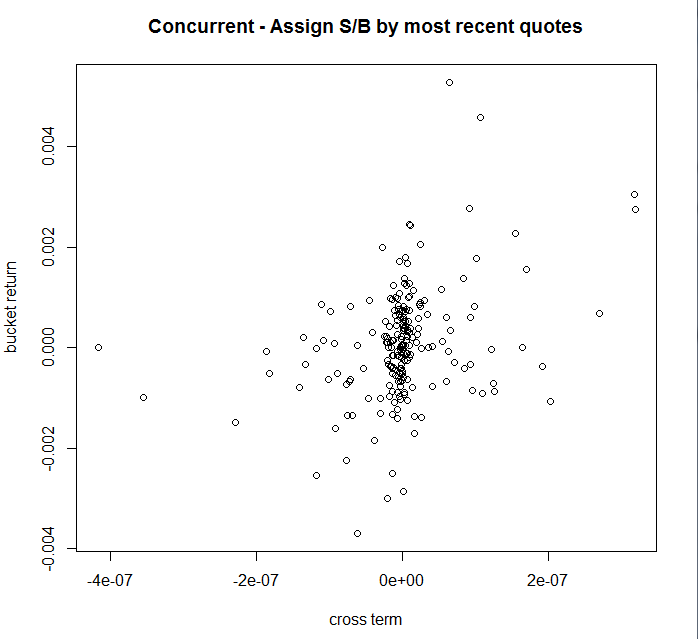
1. Use same approach as in 2 for assigning buy/sell by quotes.

|  |  |
| --- | --- |
| **BucketSize\_TimeBin\_Threshold** | **R^2** |
| 10000 \_ 60 \_ 3000 | 12.5% |
| 6000 \_ 30 \_ 1000 | 11.4% |
| 6000 \_ 150 \_ 1000 | 9.5% |
| 10000 \_ 60 \_ 5000 | 9.5% |
| 5000 \_ 60 \_ 1000 | 9.5% |
| 8000 \_ 60 \_ 5000 | 9.2% |
| 7000 \_ 60 \_ 5000 | 8.9% |
| 6000 \_ 60 \_ 1000 | 8.9% |
| 4000 \_ 60 \_ 1000 | 8.8% |
| 10000 \_ 90 \_ 5000 | 8.6% |

Best result plot as the follow:



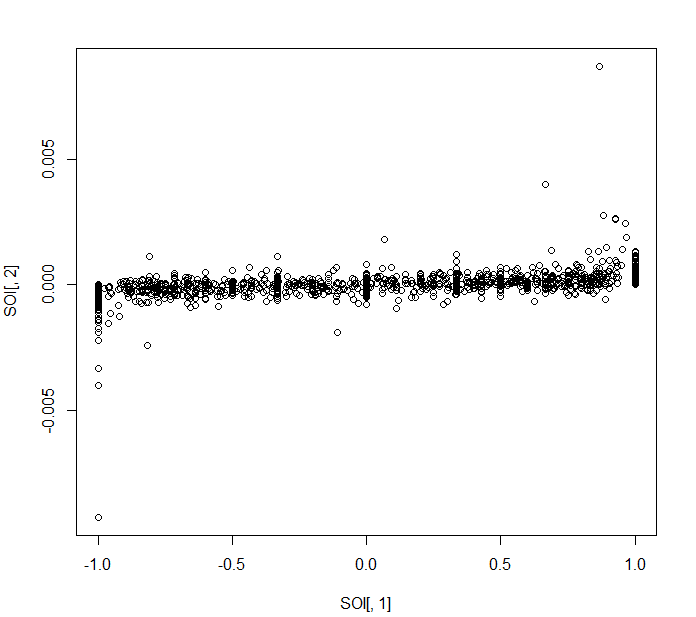
Including the cross term doesn’t help increase the adjusted R^2.



And using quotes based classification methodology produces worse results both in terms of concurrent period regression and one period ahead prediction.

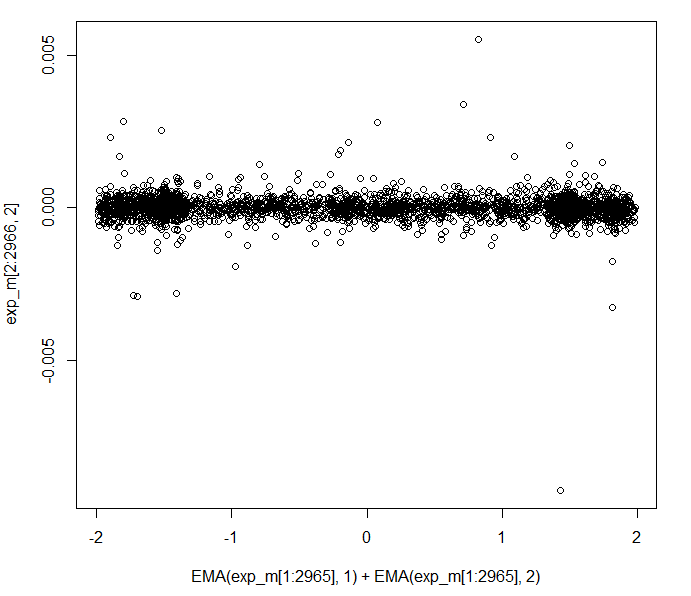
1. Time bucket concurrent analysis:

If we digress from VPIN methodology of calculating SOI, and instead define SOI using time buckets, we find the highest concurrent regression R^2 dropped to 0.233, with a time bucket size of 150s. Note when we choose the time bucket bin size, if the time bin size is large, we are unlikely obtain strong prediction results even if we obtain meaning same period association between SOI and price change.



From the graph, we can conclude that the association between SOI calculated using time buckets and concurrent period price change is not strong. And using an extended model

Using the same time bucket size of 150s, the model gives an R^2 that is virtually 0. Below is the plot:

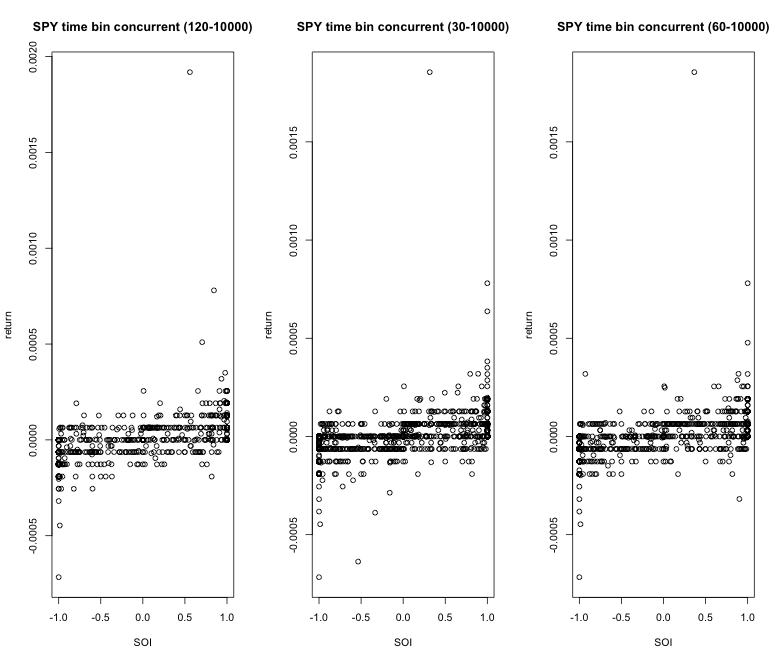


And we conclude this extended model does not have any predictive power, either.

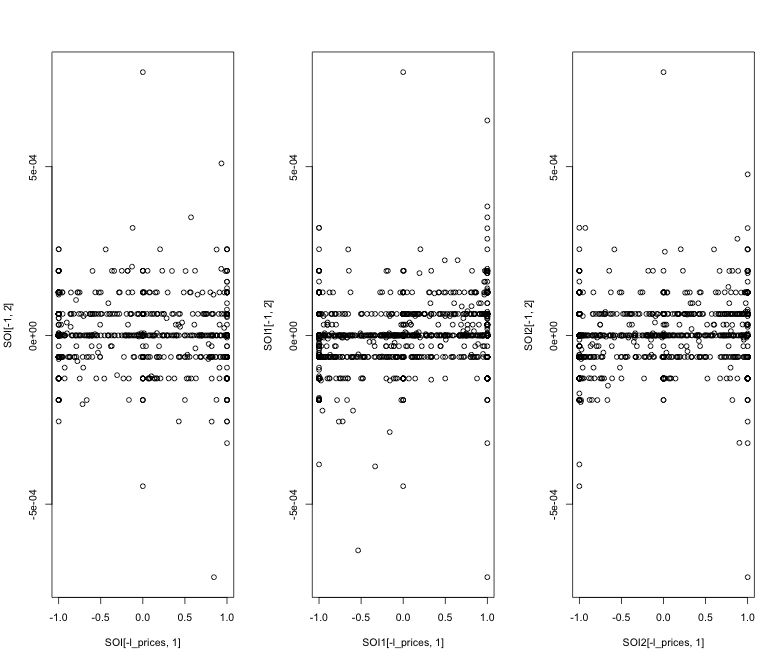
1. Time bin concurrent analysis on SPY

|  |  |
| --- | --- |
| **Time\_Threshold** | **R^2** |
| 120 \_ 10000 | 23.3% |
| 30 \_ 10000 | 23.0% |
| 60 \_ 10000 | 22.8% |
| 90 \_ 10000 | 22.8% |
| 180 \_ 10000 | 22.8% |
| 150 \_ 10000 | 21.1% |
| 30 \_ 20000 | 5.2% |
| 120 \_ 20000 | 4.7% |
| 90 \_ 20000 | 4.6% |
| 180 \_ 20000 | 4.6% |

The scatterplots of three best results:



Forecast



Appendix

Plot 1: AMZN trades size density

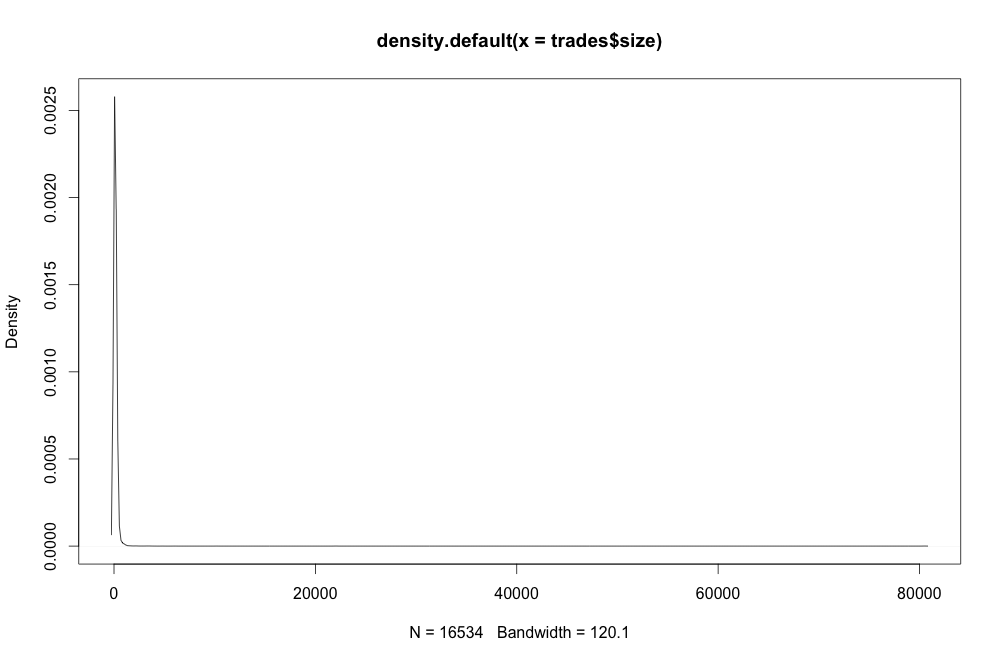


Table 1: Best results:

|  |  |
| --- | --- |
| **BucketSize\_TimeBin\_Threshold** | **R^2** |
| 10000 \_ 135 \_ 1000 | 41.30% |
| 10000 \_ 150 \_ 1000 | 41.30% |
| 10000 \_ 120 \_ 1000 | 37.40% |
| 8000 \_ 135 \_ 1000 | 37.10% |
| 9000 \_ 120 \_ 1000 | 36.60% |
| 10000 \_ 105 \_ 4000 | 36.50% |
| 8000 \_ 150 \_ 1000 | 36.20% |
| 10000 \_ 105 \_ 1000 | 36.20% |
| 8000 \_ 120 \_ 1000 | 36.10% |
| 10000 \_ 90 \_ 1000 | 35.60% |
| 8000 \_ 105 \_ 1000 | 35.10% |
| 10000 \_ 105 \_ 3000 | 35.01% |
| 10000 \_ 75 \_ 1000 | 34.95% |
| 9000 \_ 135 \_ 1000 | 34.94% |
| 10000 \_ 105 \_ 5000 | 34.13% |
| 6000 \_ 150 \_ 1000 | 33.52% |
| 7000 \_ 120 \_ 1000 | 33.14% |
| 8000 \_ 150 \_ 5000 | 33.12% |
| 6000 \_ 135 \_ 1000 | 32.86% |
| 10000 \_ 45 \_ 1000 | 32.77% |
| 10000 \_ 90 \_ 4000 | 32.72% |
| 10000 \_ 120 \_ 4000 | 32.69% |
| 9000 \_ 150 \_ 2000 | 32.63% |
| 8000 \_ 45 \_ 1000 | 32.60% |
| 8000 \_ 105 \_ 3000 | 32.56% |
| 7000 \_ 75 \_ 5000 | 32.41% |
| 9000 \_ 90 \_ 1000 | 32.39% |
| 10000 \_ 90 \_ 3000 | 32.37% |
| 8000 \_ 120 \_ 3000 | 32.20% |
| 10000 \_ 75 \_ 4000 | 32.12% |
| 9000 \_ 75 \_ 1000 | 31.98% |
| 9000 \_ 150 \_ 1000 | 31.88% |
| 10000 \_ 135 \_ 3000 | 31.82% |
| 6000 \_ 135 \_ 2000 | 31.79% |
| 7000 \_ 105 \_ 1000 | 31.54% |
| 10000 \_ 60 \_ 1000 | 31.52% |
| 6000 \_ 120 \_ 1000 | 31.44% |
| 8000 \_ 60 \_ 1000 | 31.19% |
| 10000 \_ 60 \_ 4000 | 31.01% |
| 8000 \_ 135 \_ 5000 | 30.96% |
| 9000 \_ 120 \_ 3000 | 30.80% |
| 7000 \_ 135 \_ 5000 | 30.76% |
| 9000 \_ 105 \_ 1000 | 30.76% |
| 7000 \_ 90 \_ 1000 | 30.64% |
| 8000 \_ 105 \_ 5000 | 30.60% |
| 10000 \_ 45 \_ 4000 | 30.60% |
| 8000 \_ 75 \_ 1000 | 30.59% |
| 8000 \_ 90 \_ 1000 | 30.39% |
| 8000 \_ 120 \_ 5000 | 30.38% |
| 9000 \_ 90 \_ 5000 | 30.33% |
| 10000 \_ 75 \_ 3000 | 30.30% |
| 9000 \_ 90 \_ 3000 | 30.16% |
| 10000 \_ 150 \_ 5000 | 30.10% |
| 9000 \_ 150 \_ 5000 | 30.09% |