WOODLINE
PARTNERS

Quantitative Research - Data Assignment

Report by Dev Gupta (UC Berkeley MFE)

Why do we use alternative (transaction level) data to predict sales?

Company
(full database)

→

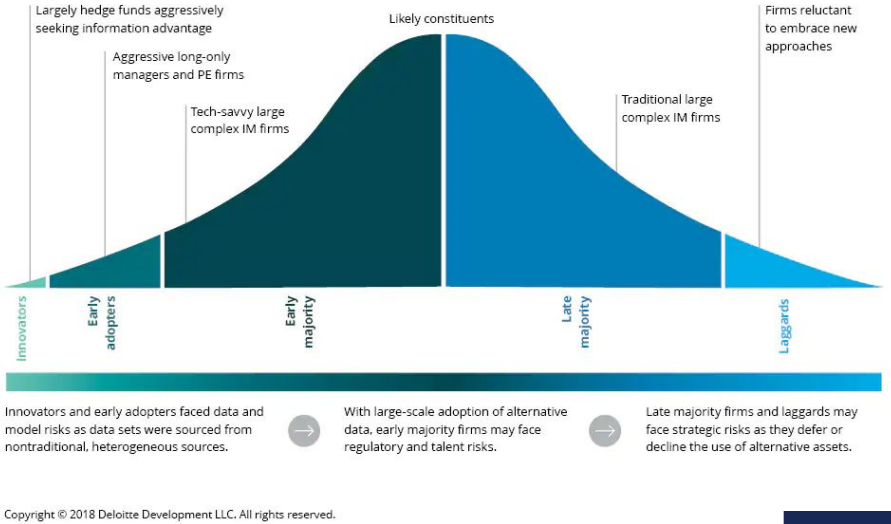
Alt. Data Vendors
(builds partial database)

→

Hedge Funds
(buys alt. data for alpha)

Alternative data is data called from non-traditional sources and used by investment firms to find a market edge.

- **Woodline Partners**, being a market neutral hedge fund, is highly dependent on generating idiosyncratic alpha from sector-specific investments.
- The goal of alternative data is to generate **incremental** (not pure) **alpha** for portfolio managers, even a narrow information advantage can improve the trading signal.
- Alternative raw data is often **expensive, incomplete, irrelevant or unverified** and the ones that could be in a good shape often fall under non-compliance issues.
- With these exercise, we try to figure out if transaction level data can be a potential indicator of actual reported sales, part of company's earnings report, that usually come out a few weeks after the end of quarter - to get an informational edge as a hedge fund.



Are there any data quality issues with this data set? If so, what did you need to clean up?

Dataset	Data Quality Issue	Issue Description	Data Cleanup
Transactions Data	**Non-Unique values**	We are supposed to have 11 different companies over 6 months, but the summary table shows that we have 110 unique companies.	The unique identification for a company is based on its letter, hence remove the rest of the words and spaces from the string. For null values in the company column, rows can be dropped.
	Duplicate & NaN values	After cleaning up for the company column, there were 37,559 duplicates in the data.	It is possible that we have greater number of duplicate rows compared to our previous unclean set as we did not have unique identifiers: use *drop_duplicates* & check for null values
	Data Outliers & Anomalies	Item_price seems to have a few outlier data points (max: 999999)	Using IQR, we replace outliers with a NULL value. We could have imputed the missing values with Mean but there were only 20 values - hence it would not make a big difference in our dataset.
Transactions & Report Data	**Incorrect date formatting**	Date is not in the correct datetime format	Convert column to datetime via pandas and check min/max are starting from Q3 & ending at Q4

Are there any companies for which this data set would have been a good predictor?

Prediction Accuracy = Predicted Sales(Total Transaction Sales) /Actual Reported Sales

Prediction Accuracy

	company	prediction_accuracy
0	A	0.107
5	F	0.096
8	I	0.090
6	G	0.077
2	C	0.076
9	J	0.060

From a very preliminary analysis, it would seem like the transaction data is a good predictor for company A, F, and I, but we also need to note the following bias: Actual reported sales for these firms are way lower compared to the other firms, which means a lower denominator, and hence improving the accuracy value *(same txn data size given for all firms)*

Trend Prediction

Company E & F clustered together would typically not improve heavily in their sales value from Q3 to Q4, so if we had Q3 data, we can predict Q4 sales to be almost same.

Company H would most likely be having a heavy jump in sales from Q3 to Q4, and reported sales highly advocates this spike in sales for H.

Correlation Prediction

Please check correlation matrix for more information.

All company sales (A-K) except company F & company I are highly positively correlated to each other. Market-neutral (like Woodline's) strategies involve long and short positions in two different securities with a positive correlation. **A pairs trade strategy** is based on the historical correlation of two securities.

Sales of **Company I are highly negatively correlated** to all the other companies (except F). From trading point of view the -ve correlations can be used to **diversify the portfolio and reduce overall risk.**

Are there any companies for which this data set would have been a bad predictor?

Prediction Accuracy = Predicted Sales(Total Transaction Sales) /Actual Reported Sales

Prediction Accuracy

4	E	0.057
1	B	0.056
3	D	0.055
7	H	0.054
10	K	0.048

Company H, D, K, and B (bad predictors) on an average have the highest quarterly sales amongst the eleven unique firms whereas from the alternative data's prediction accuracy, it is clear that the accuracy for these firms would be lower as they have much larger reported sales compared to the other firms.

Trend Prediction

The data is quite stationary within individual quarter time frames but there is definitely a jump in sales from Q3 to Q4, could be a **seasonality component**.

Company A, C, and G would most likely be having a heavy jump in sales from Q3 to Q4, but reported sales show stability across both quarters.

Correlation Prediction

Please check correlation matrix for more information.

Company F has very low correlation to all the other companies. It could be utilized for a diversification trading strategy, but from a sales prediction behaviour – hypothetically if we only had sales of company F from alternative data vendor, we cannot predict/extrapolate the sales of other companies. Additionally, if there's no discernible correlation between two underlyings, there's no way to enter into a pairs trade involving the two with a high degree of confidence in the potential outcome.

Are there any potential issues that could bias your analysis?

It could be the case that the **alternative data vendor has more bias towards data for specific companies as they are probably more involved with these user transactions** *(example 's A's users are mainly transacting with the data vendor's platform)*

Interestingly, all of the companies are only selling 5 products, this again shows the l**ack of representation of the total product assortment that the companies might have** but does not come as part of the intersection between alternative data vendor and the hedge fund. There could also be a bias as potentially the data vendor has been asked to share the sales of 5 specific products across the 11 unique companies.

In general there could be a lot of selection bias in terms of data, for example the **user_ids provided to us are only of that population which is transacting online**, the ones which are not might represent a huge popn. of reported sales.

Although, highly unlikely without the full knowledge of transaction data dictionary, **the sales price might be reported at a per $1,000 level (quite common in alt. data)**- which sort of proves that our data is highly predictive for A, F, I, & G

No firm would give their data freely to a buy-side firm and these data points usually come from a mediator, and as the number of middle-layers between the buy-side firm's data acquisition process and the actual company data increases, the prediction accuracy goes down.

Are there any companies that are more popular than others? How did you measure “popularity”?

Popularity can be measured by: Total Users + Customer Frequency + Per-user spend + Total transactions
We can also check if users are ordering from multiple companies in the transactions dataset (check scatter plot below)

Popularity Index	Definition (why?)	Top 5 Ranking
#unique users	Metrics around customer database are the best way to showcase the popularity of firms	H > K > B > C > G
Customer Frequency	Number of transactions/Number of Unique Users	K > B > D > F > I
User spending	Total Sales Value / Number of Unique Users	**A > C > D > H > F** *(a different ranking)*
Total transactions	Count of user_ids from the transactions dataframe	K > B > H > C > I

Trying to cluster user groups based on multiple metrics consolidated from the original dataset (daily data)