**Capstone Milestone Report**

1. **Goals:**

We are attempting to predict the probability that a participant in our survey will require human contact to remind them to submit. We can then anticipate which participants will need more attention and tailor follow-up (i.e. emails, phone calls, etc.) to them, that is customized based on the model results, before personal contact. This follow-up will not only reduce the amount of contact needed over time and save human resources for the administrator of the survey, but it should improve the participation in our survey. The model can also be adapted to other problems the client has with retention, such as membership retention and trade show attendance retention. The clients are the administrators of the survey.

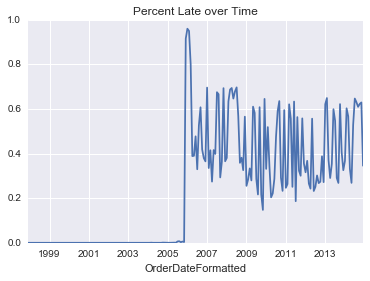
**II. Data Cleaning and Feature Engineering:**

The data set is very rich and contains many variables, though we have decided to use only the most pertinent and merged them with a few outside data sources. The most important ones which will be included in the analysis are:

1. Company ID - a unique identifier for each company
2. Order Date - the applicable month for each set of data
3. Total Sale value of the current month
4. Units Sold in the current month
5. Days Late they submitted in the current month
6. S&P 500 open price of the current month
7. Industrial production value of the current month
8. Received first reminder email dummy variable
9. Received second reminder email dummy variable
10. Received personalized call (target) dummy variable
11. Dummy variable for if company backfilled data
12. Average of company’s last three month’s sales
13. Average of company’s last three month’s days late
14. Last month’s days late
15. Last month’s average machine value
16. Last month’s total sale value
17. Received personalized call last month dummy variable

The limitations are that we do not have the ideal number of features for prediction; feature selection has only identified 3 as highly significant.

Cleaning required removing the former half of the dataset because it was found to conform to different standards than the latter half, as shown below. Data from before May 2006, which is when the spike occurs, is excluded.



Days Late was being measured in an inconsistent format across the two halves. Therefore meaningful comparisons could not be made across the two halves. We also needed to tack two external datasets, the S&P 500 values and industrial production values (features 5-6), and do some feature engineering, creating the latter ten variables in the above list (features 7-17). In addition, we needed to do some formatting to get the dates in the correct format. Finally and perhaps most significantly, we tossed out all observations where days late was greater than 90 days, as the client has a rule that participants are ejected from the survey after 90 days. Observations where days late is greater than 90 days are therefore only because they are backfilled from a much later point in time, skewing the data set. This was a small minority of the dataset (12501/392068, or 3.1% of records).

External datasets include the S&P 500 values and industrial production. The reason for this inclusion is because we had a hunch that poor business conditions would lead to de-prioritization of the survey amongst our participants.

**III. Exploratory Data Analysis:**

EDA yielded significant insights about the dataset. Several of the most prominent graphs are displayed below with explanation of consequences.

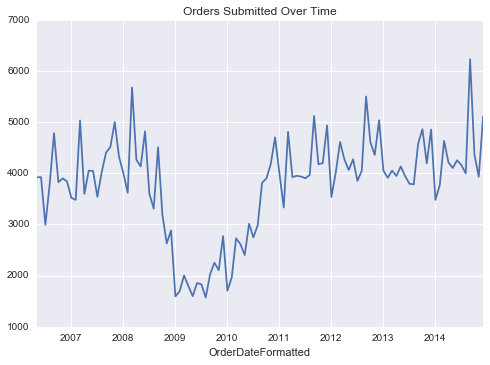


Exhibit 1: Number of total orders submitted over time, mirroring the S&P 500 and industrial production index. Correlation confirms hypothesis that submissions are influenced by domestic business conditions.

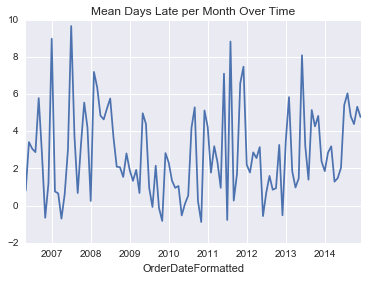


Exhibit 2: Mean days late over time appears to stay in fairly consistent range over time, indicating no significant exogenous shocks. This indicates that our target does not appear to be heavily influenced by any major events, whether external to the client or internal

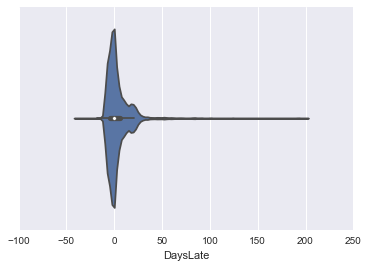


Exhibit 3: Violin plot of days late indicates there is not an epidemic of late submissions. Rather, most participants submit early or on time. This indicates that our target is not too broad and is a healthy target. This is further confirmed by the analysis below - showing that overall all time periods in our dataset, about 23% of companies need personal contact (IsCall indicates personal contact):

**https://lh4.googleusercontent.com/_8--C63dfsBdIdRoWP4AeOqH3OPHhZPVrM_V-BtsTz5tkdzvDQP97lSbHjJ45d4bzGXMEGqhdYzyYHmBV1rkGEqkOD00IyhgG9VFzf-0hYqfgYIsFTgtbTq2h4xcG6FjGDwvpO3O**

Below is a correlation matrix which shows our features’ correlation to each other, and to the target. Three\_MMA\_Late, DaysLateLastThree, DaysLateLast, IP, and IsCallLast have the highest correlation to the target.

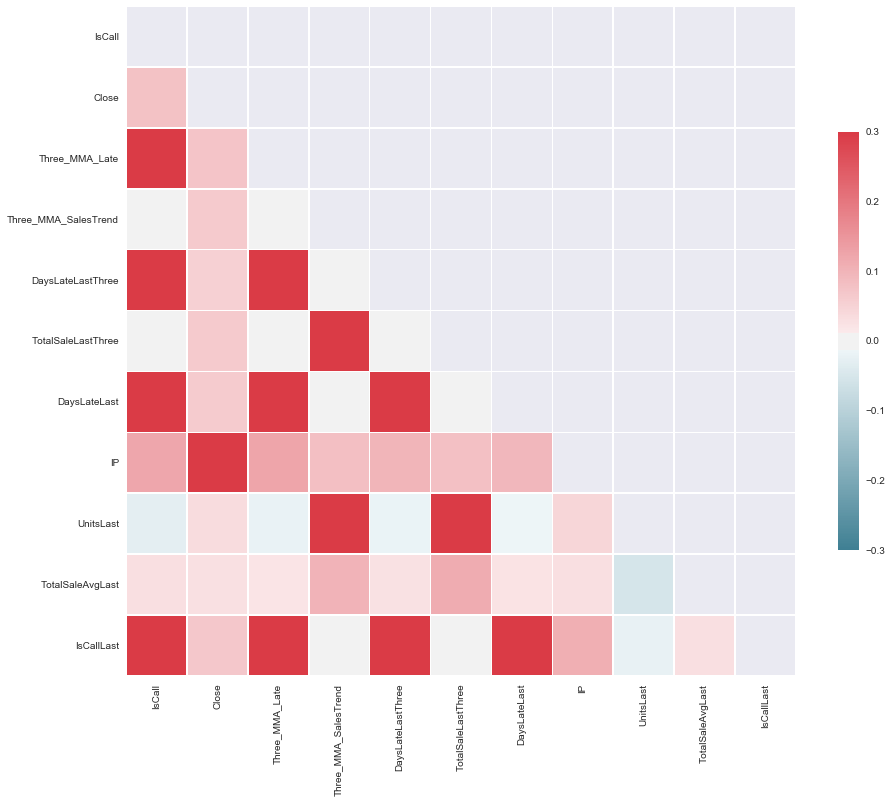


Exhibit 4: Heat map of the significance of each feature. Note that several features have high correlations to each other. However, multicollinearity is smoothed over by our models, as the most significant will be selected and the others de-prioritized.

**IV. Next Steps:**

Because of the limited number of features, we have chosen not to narrow the feature set down. Rather, we will use all meaningful features (that is, we excluded all features which cannot be used for prediction of future states, like the current month’s units sold - which cannot be used to predict current month’s submission status). In other words, we excluded features that we cannot obtain until the month of the target prediction. The final feature set (10 features) used is below:

      'Close', 'Three\_MMA\_Late', 'Three\_MMA\_SalesTrend',  
      'DaysLateLastThree', 'TotalSaleLastThree', 'DaysLateLast', 'IP',  
      'UnitsLast', 'TotalSaleAvgLast', 'IsCallLast'

This will be modeled as a classification problem using logistic regression and we will use applicable machine learning methods to arrive at a “score” of whether or not they will require a personalized call. Two possible models to explore include random forest and kNN. We will then perform cross validation to evaluate the efficacy of our models.

This does not differ from our original approach, though we are changing our target to be “needs personal contact” vs. “did not submit”, as the former historically occurs 23% of the time while the latter occurs 3-5%, a target that is small enough to create unnecessary complications with our model. The business impact of discovering effective ways to classify either is identical, as the consequence of both will be the same - a personalized and pre-emptive form of contact encouraging participants to make a submission.

The next step will be to experiment with, test the accuracy of, and implement machine learning models most appropriate for our feature set and problem. Normalizing and/or regularizing the data to fit these models may also be required. Additional feature engineering and external data sets may be included to fine tune the accuracy of the model.