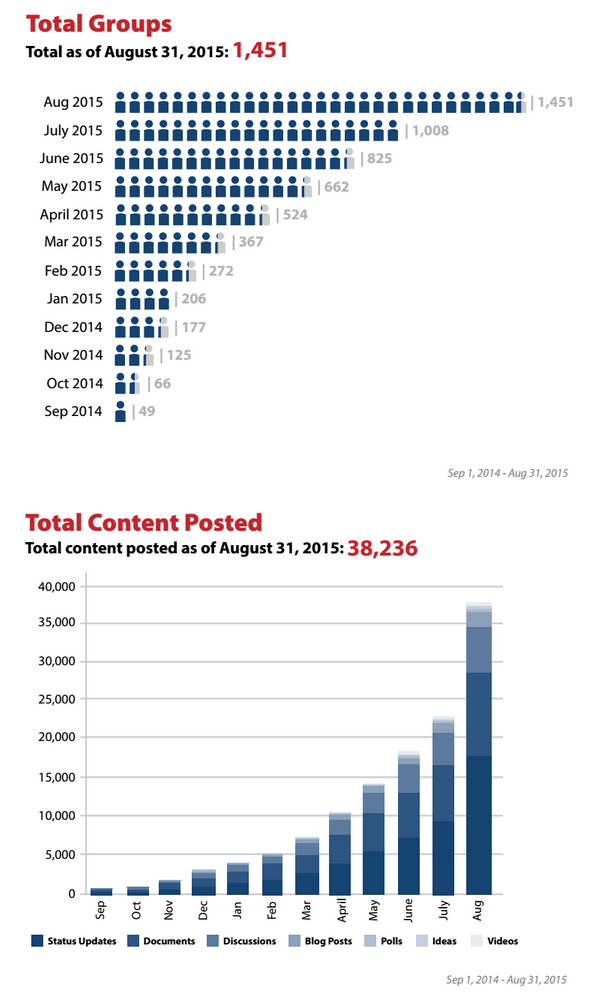
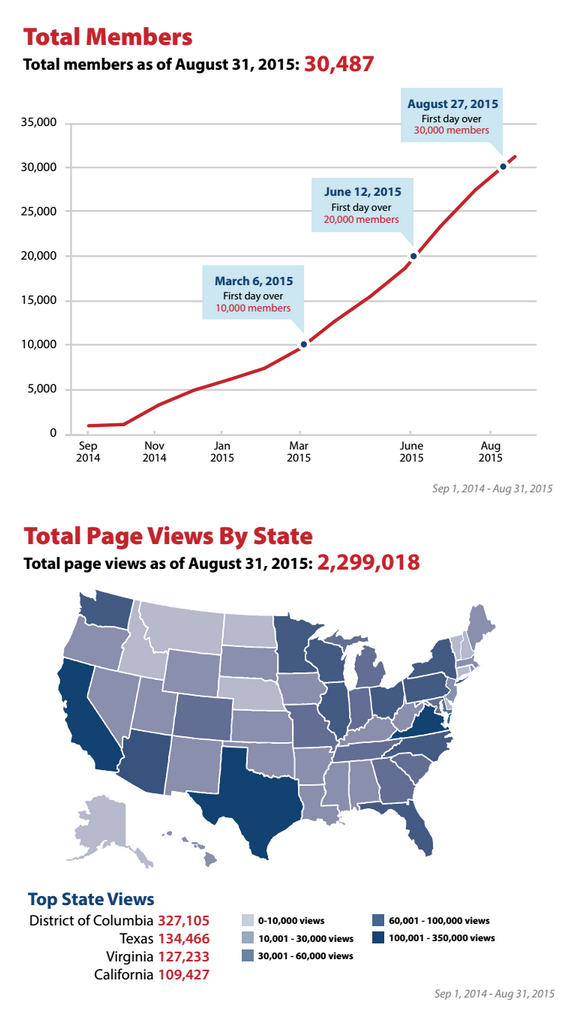
Successful

Enterprise Social Collaboration

Using Data Science to Solve the Problem of Generating Successful Adoption

# Micah Markman

# DAT8 - Fall 2015 - Final Report

Successful

Enterprise Social Collaboration

Using Data Science to Solve the Problem of Generating Successful Adoption

# The Challenge

Enterprise Social Collaboration blew up as technology that could revolutionize the knowledge workplace in the last 3 years. A host of pure-play start-ups brought the simplicity and usability of Facebook and LinkedIn into the workplace in an effort to revitalize interest collaboration software and deliver the value promised by the intranets and corporate knowledge portals that originally came into being in the mid-to-late 2000’s. These companies promised that replicating the experience of Facebook in the enterprise would generate the user adoption that is critical to actually deriving the return on investment these solutions can deliver. Despite these promises, many implementations have suffered the same adoption challenges seen in the prior generation of intranet projects. My theory is that these implementation are still failing to execute on key organizational change management best-practices; specifically failing to find ways to bring users into the system and demonstrate the value in a way that causes minimal pain to each user and may even be fun.

How can data science help solve this problem of getting users to adopt these social collaboration platforms and support the overall user behavior change process? Up until now, all of the efforts that I’ve seen to support and drive changing user behavior has been done based on anecdotal evidence and historical best practices. For example, we need to do lots of user training sessions because that’s what is done. Another example that has taken off is “gamifying” usage of these systems by rewarding users with points for completing a variety of usage activities within the system. The theory is that if you can get a user to develop certain habits when they first start using the system, the will continue even after the user tires of “game”. However, the rules that dictate how users earn points have been generated based on anecdotal evidence and theories about what will work. My belief is that we can use data science to understand the actual behaviors and attributes of users who become successful drivers of value and the create materials and functionality in the platform to drive all users follow the path laid out by early successes. Specifically, I believe that if we can train a model to predict whether a user will “adopt” the platform, then we look at what the model believes impacts this classification and use this data to make more targeted interventions with users as well as better “games” to build formative habits to get more users to follow the path of the successful adoption.

# Focusing the Problem and Hypothesis

# Source Data

The core of the data available for this lives in a couple of tables in a couple of SQL databases:

User

| Column | Data Type | Notes / Description |
| --- | --- | --- |
| userid | bigint | unique id for user |
| creationdate | bigint | easily convertible to timestamp for readability |
| userenabled | bool | 0 no longer enabled  1 enabled |
| usertype | bool | real user (0) vs. placeholder for invited user (1) |

User Profile

| Column | Data Type | Notes / Description |
| --- | --- | --- |
| userid | bigint | reference to user.userid |
| fieldid | bigint | which profile field this row is for (reference to profilefield |
| levelid | bigint | level of sharing of profile data element (reference to profileseclvl |
| value | varchar | actual user profile field value |

User Activity

| Column | Data Type | Notes / Description |
| --- | --- | --- |
| userid | bigint | reference to user.userid |
| activty\_type | varchar | type of activity |
| activity\_ts | datetime | timestamp of activity |
| metadata | various other | other information about the activity (ignored for now, this turned out to be very confusing and requires more investigation) |

# Pre-processing and Scrubbing

1. Limit to users have joined system between October 2014 and April 2015; sufficient count and this way any other variables outside data in system can be controlled for (calendar cycles, other activity within organization, etc).
2. Normalizing data to show the following per user (big effort to get the sql worked out). I tried getting to this in using Pandas data tables but things were really slow and I kept having to tweak things; much easier/faster for me to do in SQL:

User Profile

* + UserId
  + Title (free text self-description)
  + District (general region, 5 values after cleaning)
  + Location (fixed list after some cleaning but more than 200 unique values)
  + Occupation (fixed list after some cleaning with about 50 unique values)
  + Phone provided (bool)
  + Biography provided (bool)

User Activity

* + User Id
  + Activity Type (~40 types)
  + Month 1 Count
  + Month 6 Count

Read in these as CSV files and then used Pandas to Pivot the Activity data into columns and join the columns together.

Then factorized the two bool features and one-hot encoded the categorical features.

Resulted in approximately 400 features.

# Data Exploration

1. ~48% of users drop to 0 activity after 6 months
   1. This includes users who leave the organization completely so this “end-state” is over-reported; trying to find a way to clean these users out of data.
2. ~21% of users are “Contributors” after 6 months

Intuitions based on playing with data

1. Thus far, no clear pattern in first month personal activity; some users do a lot and then go quiet; others start with little activity and then go nuts. This lead to re-definition of activity into “Small/Large”; haven’t had a chance to really understand this.
2. Membership in an “Active” group that adopts tool seems to correlate with continued participation; do we need to further engineer this feature? A significant percentage of “Contributors” are members of top 50 groups by activity.
3. Anecdotal evidence / theory that number of relationships built with active users leads to activity (looking to engineer a feature to represent this).

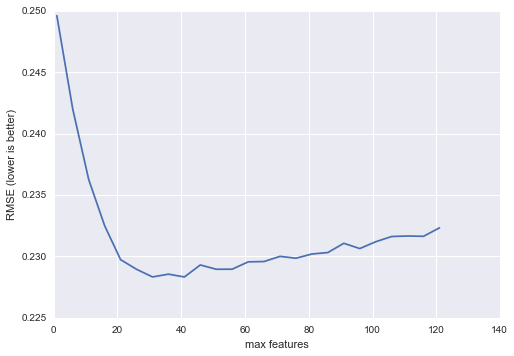
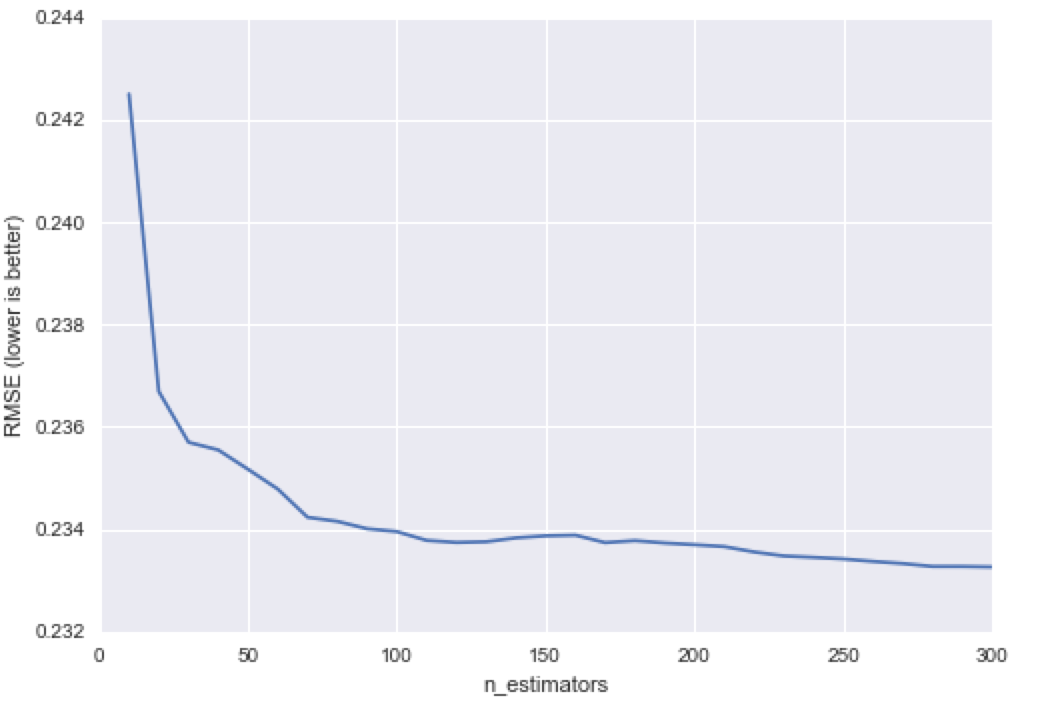
# Feature Selection and Data Transformation

# Modeling Process

Early in my process, I tested logistic regression and KNN and Logistic Regression but I had trouble having a good intuition about features and really things didn’t do that much better than the model. Honestly, at this point, I was in data gathering mode and was using Logistic Regression just as a sanity check against my data with goal of moving to Random Forests.

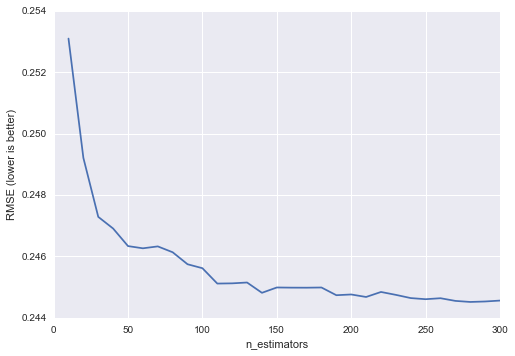
## Random Forests:

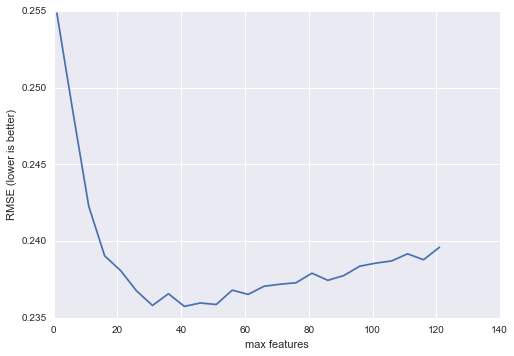
First optimized on predicting consumers response.

Ran through Random Forests and optimized for n\_estimators and max\_features. Here are the two graphs that lead me to tuning with n\_estimators of 100 and max features of 40:

Resulted in Out of Bag Score of 0.6682882775449365 and using train test splice got an accuracy of 92.8% and AUC of 0.962 and following confusion matrix:

|  |  |
| --- | --- |
| 2470 | 89 |
| 138 | 497 |

Now optimized on predicting contributors response. Here are the two graphs that lead me to tuning with n\_estimators of 150 and max features of 35:



# Challenges and Successes

1. Scrubbing training attendance data has proven to be more work than currently have time for; have to eliminate this feature but want to bring back.
2. Realized that there were a lot of users that joined but didn’t really ever actively engage and fell off very quickly. These seem to introduce a lot of noise/bias into the system. May want to rerun some models after excluding these users.

# Conclusions

I was able to predict with a bit more accuracy that the null model but

# Immediate Next Steps

* Look into doing better regularization, ran out of time and I think this is having an outsize impact on decision tree visual.
* A new month of data is about to be available: opportunity to predict with truly out of sample data and further refine model
* Add in features for social data, membership in key groups, associations with key users (after investigating this; flaws in pulling data out of SQL were discovered).
* Update model to predicting whether a user is going to drop-out based on current rolling months worth of activity; enables team to re-target users before they drop
* Rework overall system to bucket data by week as opposed to by month.

# Business Applicability

Cool