

A dark blue graduation cap with a black tassel and a rolled white diploma tied with a red ribbon are positioned in the upper left corner of the slide.

iMentor Project

...

December 15, 2016
Team Awesome



iMentor Background

Who?

iMentor is a non-profit organization that facilitates mentoring relationships to ensure more students from low-income communities enroll and graduate from college.

Where?

Serves more than 6,000 students through its direct-service programs in New York City, Chicago, and the Bay Area

How?

By partnering with public high schools, the program matches students with mentors for the majority of their high school career.

- Weekly messages and chat conversation
- Monthly in-school meetings and many more!

Problem Statement:

Mentor persistence

With the ultimate goal of college completion, one of the iMentor's Core Metrics of quality is Mentors under the program execution rubric. A mentor dropping out of the program before its full length could impact the mentee's success in the program.



Understanding the problem

High dropout in year 2-3

iMentor's own analysis revealed that a significant percentage of mentors do not complete the 3- or 4-year mentoring commitment, with most dropping out over the second and third year.

2 Program Types

- College Ready (4-year program)
- College Transition (3-year program)

Any indicators?

- Mentors demographic attributes
- Mentors behavior: message frequencies, responsiveness and length

Project objective:

Establish a successful framework
for repeatable and continued
analysis of possible indicators of
mentors dropout



In scope

The analysis will cover:

- Mentors demographic information
- Message log data for one school year 15-16 i.e. excludes the content of messages
- Mentors participating in iMentor NYC only

Note 1: Additional datasets are still processed in our ETL since will be used in the future analysis.

Note 2: Further data is recommended to be included as a next step of this project.



1. Analysis Framework

Framework Chart

Data sources

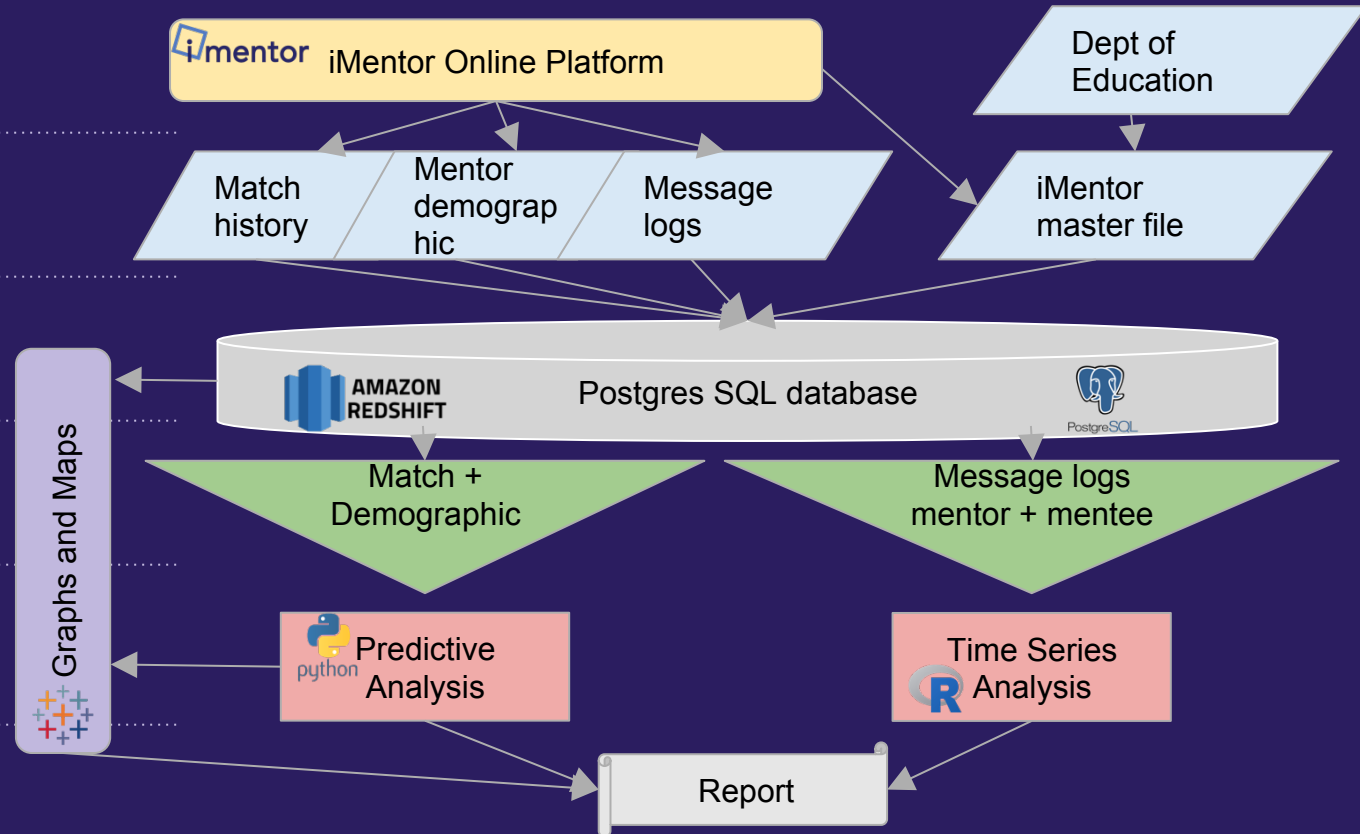
Data files

Data ingestion and transformation

Data extraction and merge

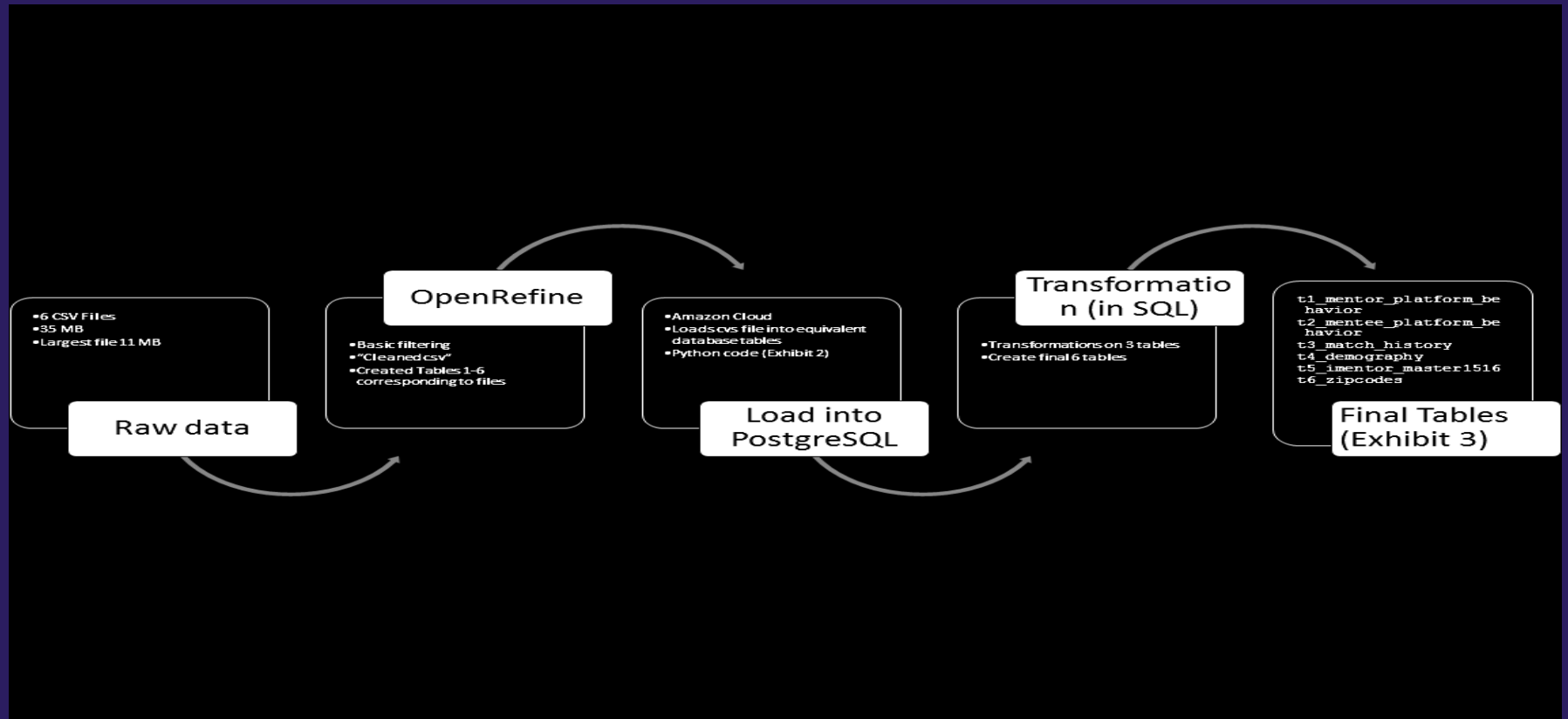
Additional transformation and analysis

Report and Visualization



2. Data Processing

Extract - Load - Transform



Our Database on Amazon

Snipping Tool

File Edit Tools Help

New Delay

RDS Dashboard

- Instances
- Clusters
- Reserved Purchases
- Snapshots
- Security Groups
- Parameter Groups
- External Licenses
- Option Groups
- Subnet Groups
- Events
- Event Subscriptions
- Notifications

Launch DB Instance Show Monitoring Instance Actions

Filter: All Instances Search DB Instances... Viewing 1 of 1 DB Instances

Engine	DB Instance	Status	CPU	Current Activity	Maintenance	Class	VPC	Multi-AZ	Replica
PostgreSQL	projectawesome	available	3.39%	0 Connections	None	db.m3.medium	vpc-5d5b143a	No	

Endpoint: projectawesome.cxbeoyk1nuf.us-east-1.rds.amazonaws.com:5432 (authorized)

Configuration Details		Security and Network		Instance and IOPS	
ARN	arn:aws:rds:us-east-1:730936178929:db:projectawesom	Availability Zone	us-east-1e	Instance Class	db.m3.medium
Engine	PostgreSQL 9.5.4	VPC	vpc-5d5b143a	Storage Type	General Purpose (SSD)
License Model	Postgresql License	Subnet Group	default (Complete)	IOPS	disabled
Created Time	November 30, 2016 at 5:30:40 PM UTC+8	Subnets	subnet-aa5a41f2 subnet-5b140e71 subnet-da5fbae6 subnet-3d25ef74	Storage	5 GB
DB Name	awesome	Security Groups	default (sg-11e55d6b) (active)		
Username	awesome_admin	Publicly Accessible	Yes		
Option Group	default:postgres-9.5 (in-sync)	Endpoint	projectawesome.cxbeoyk1nuf.us-east-1.rds.amazonaws.com		
Parameter Group	default.postgres9.5 (in-sync)	Port	5432		
Copy Tags To Snapshots	No	Certificate Authority	rds-ca-2015 (Mar 5, 2020)		
Resource ID	db-XYV3PKLHOY5X4J7VF3ECRVRBW				
Encryption Details		Availability and Durability		Maintenance Details	
Encryption Enabled	No	DB Instance Status	available	Auto Minor Version Upgrade	Yes
		Multi AZ	No	Maintenance Window	wed:04:46-wed:05:16
		Automated Backups	Enabled (7 Days)	Backup Window	08:53-09:23
		Latest Restore Time	December 14, 2016 at 8:53:02 PM UTC+8	Pending Maintenance	None

Data Processing App (10 sub-programs)

```
levi@levi-VirtualBox: ~/MIDS/Project
4 dnocror-x 3 levi levi 4096 Dec 14 17:33 .
4 dnocror-x 4 levi levi 4096 Dec 1 11:17 ..
42980 -m-n-e-r-- 1 levi levi 44011520 Dec 14 17:32 backup.tar
4 dnocror-x 2 levi levi 4096 Dec 5 22:21 ipynb_checkpoints
8 -m-n-e-r-- 1 levi levi 4505 Dec 12 21:36 post_create_tables.py
12 -m-n-e-r-- 1 levi levi 10027 Dec 11 21:19 post_load_master_table.py
12 -m-n-e-r-- 1 levi levi 9173 Dec 1 14:35 post_load_master_table.sql
4 -m-n-e-r-- 1 levi levi 898 Dec 11 21:19 post_load_match_history.py
4 -m-n-e-r-- 1 levi levi 422 Dec 1 14:53 post_load_match_history.sql
4 -m-n-e-r-- 1 levi levi 875 Dec 12 12:04 post_load_mentee_behavior.py
4 -m-n-e-r-- 1 levi levi 875 Dec 11 21:19 post_load_mentor_behavior.py
4 -m-n-e-r-- 1 levi levi 1114 Dec 11 22:51 post_load_mentor_demo_table.py
4 -m-n-e-r-- 1 levi levi 876 Dec 11 23:18 post_load_platform_behavior.py
8 -m-n-e-r-- 1 levi levi 6846 Dec 12 21:21 post_process_master_table.py
4 -m-n-e-r-- 1 levi levi 2111 Dec 11 23:27 post_process_match_history_table.py
12 -m-n-e-r-- 1 levi levi 12274 Dec 12 11:37 post_process_mentor_demo_table.py
896 -m-n-e-r-- 1 levi levi 914333 Dec 4 14:24 Table_1_match_history.csv
5480 -m-n-e-r-- 1 levi levi 5607959 Dec 11 22:54 Table_2_mentor_demo.csv
9712 -m-n-e-r-- 1 levi levi 9943775 Dec 4 14:24 Table_3_mentor_behavior.csv
7924 -m-n-e-r-- 1 levi levi 8113946 Dec 11 23:10 Table_4_platform_behavior.csv
9480 -m-n-e-r-- 1 levi levi 9705326 Dec 5 14:08 Table_5_mentee_behavior.csv
9436 -m-n-e-r-- 1 levi levi 3659943 Dec 5 14:07 Table_6_master_eoy.csv
leviglevi-VirtualBox: ~/MIDS/Project$
leviglevi-VirtualBox: ~/MIDS/Project$
```

```
#!/usr/bin/perl
# Program to create Table_1_match_history.csv from Table_1_match_history.sql
# Usage: perl create_table_1_match_history.pl <table_name>
# Example: perl create_table_1_match_history.pl Table_1_match_history

my $table_name = "Table_1_match_history";
my $table_name = $ARGV[0];
my $table_name = "Table_1_match_history" if ($table_name eq "");

my $table_name = "Table_1_match_history";
my $table_name = $ARGV[0];
my $table_name = "Table_1_match_history" if ($table_name eq "");
```

```
#!/usr/bin/perl
# Program to create Table_2_mentor_demo.csv from Table_2_mentor_demo.sql
# Usage: perl create_table_2_mentor_demo.pl <table_name>
# Example: perl create_table_2_mentor_demo.pl Table_2_mentor_demo

my $table_name = "Table_2_mentor_demo";
my $table_name = $ARGV[0];
my $table_name = "Table_2_mentor_demo" if ($table_name eq "");

my $table_name = "Table_2_mentor_demo";
my $table_name = $ARGV[0];
my $table_name = "Table_2_mentor_demo" if ($table_name eq "");
```

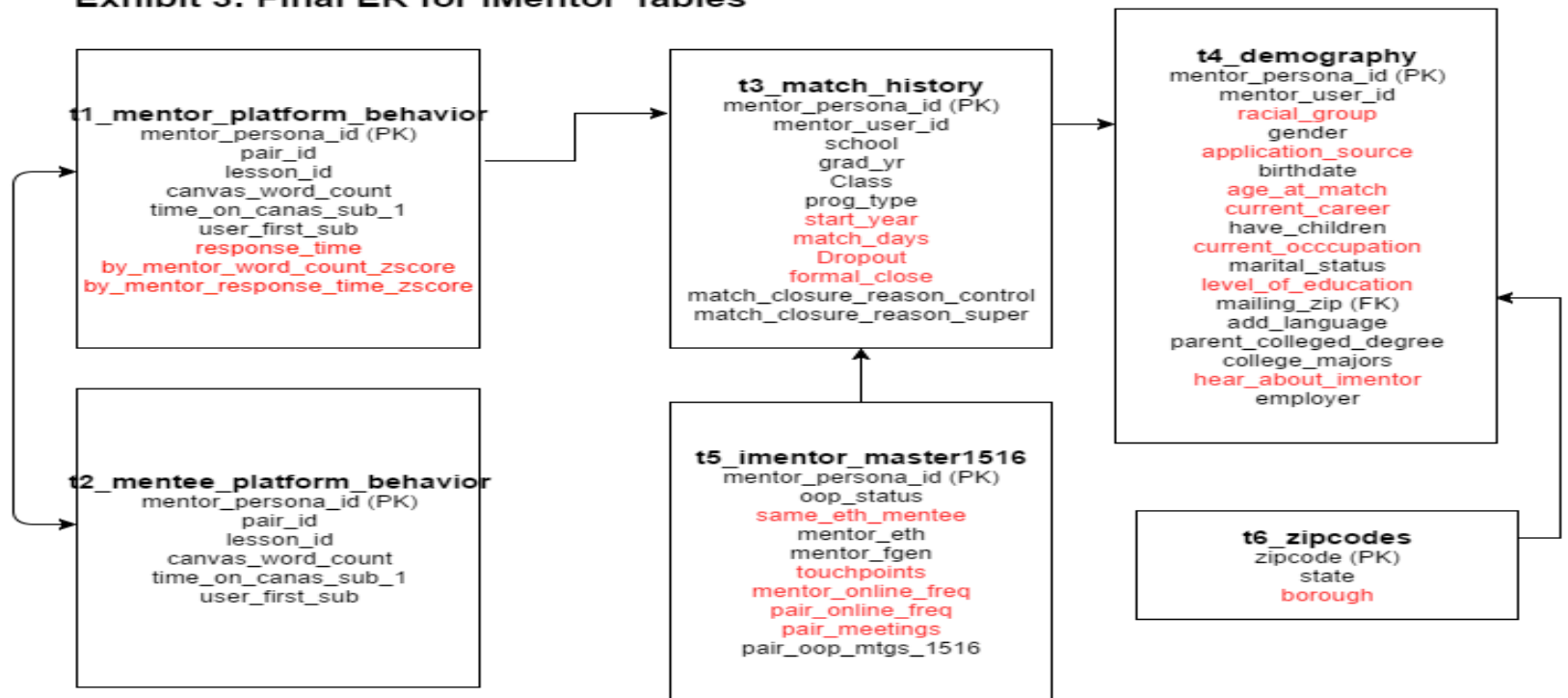
```
#!/usr/bin/perl
# Program to create Table_3_mentor_behavior.csv from Table_3_mentor_behavior.sql
# Usage: perl create_table_3_mentor_behavior.pl <table_name>
# Example: perl create_table_3_mentor_behavior.pl Table_3_mentor_behavior

my $table_name = "Table_3_mentor_behavior";
my $table_name = $ARGV[0];
my $table_name = "Table_3_mentor_behavior" if ($table_name eq "");

my $table_name = "Table_3_mentor_behavior";
my $table_name = $ARGV[0];
my $table_name = "Table_3_mentor_behavior" if ($table_name eq "");
```

Schema (Final Tables)

Exhibit 3: Final ER for iMentor Tables



3. Mentor demographic data Predictive Analysis

Key Findings

- Formal closure and dropout mentors group size significantly differ by program type
- Demographic attributes are not sufficient alone to accurately predict probability of whether a mentor will drop out

Program Types

Figure 1 - entire match history

4-yr program mentor dropout > formal closure

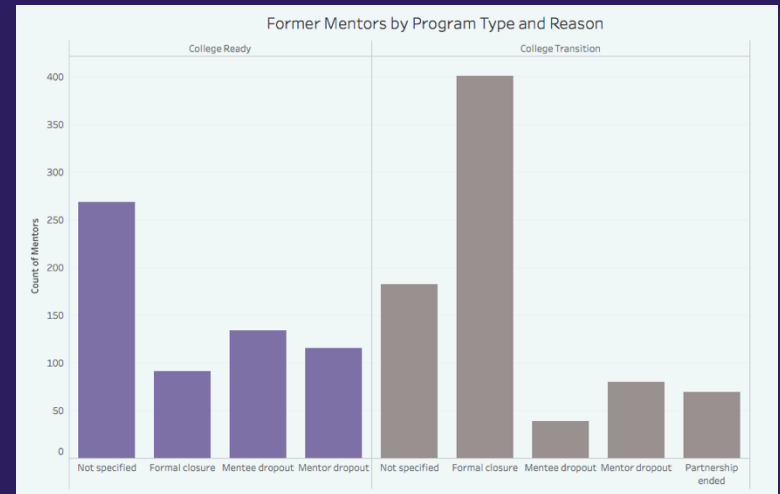
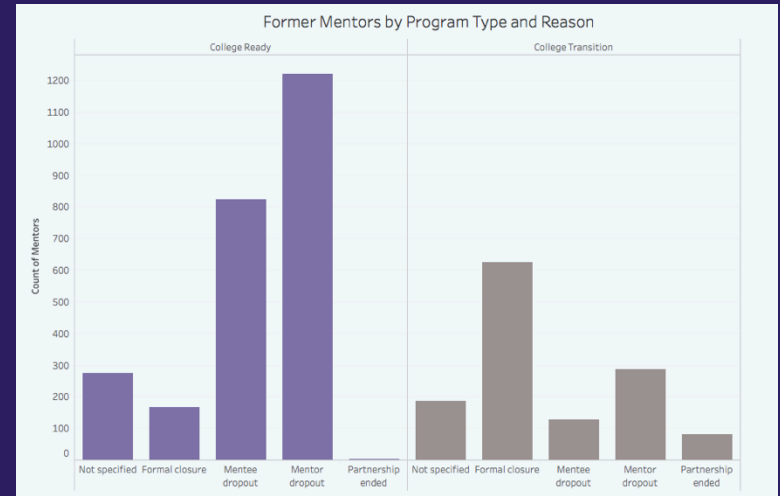
3- yr program mentor dropout < formal closure

Figure 2 - 2012 match start year

4-yr program mentor dropout > formal closure

3- yr program mentor dropout < formal closure

Observation: There are more mentor dropouts than formal closure for 4-yr program and the opposite case for 3-yr program. However in case of 4-yr program, such difference narrows for 2012 match start year cohort which covers a period of at least 4 years to 2016 data collection year.



Question:

Can mentors demographics attributes be indicators of mentor dropouts?

Method:

Predictive analysis by training mentors demographic attributes in a classification model



Mentors Demographic Attributes

List of attributes in the predictive analysis of dropouts:

- Age at match
- Gender
- Marital status
- Racial group
- Career
- Level of education
- Parent college degree
- Have children
- How heard about iMentor

Note: some variables which contain overlapping characteristics (e.g. occupation vs. career) have been removed.

Figure A. Dropout and Demographic features correlation matrix

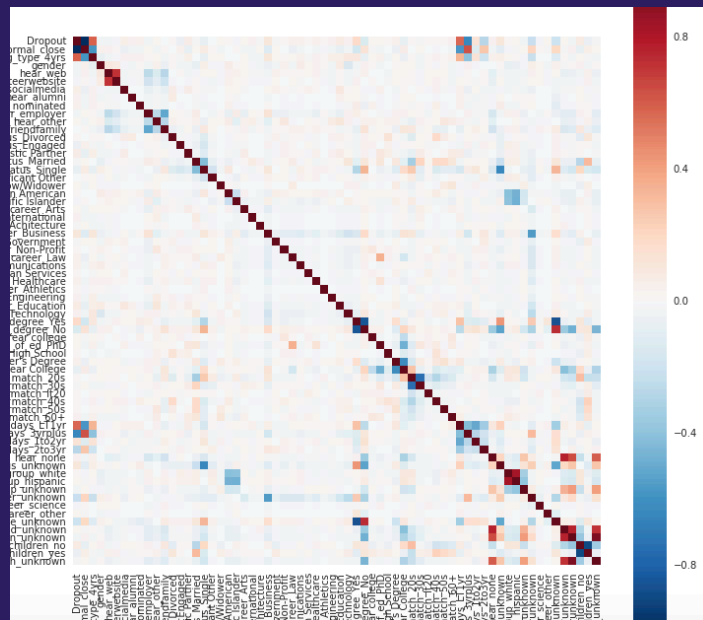


Figure A shows that not a lot of demographic attributes of mentors have a strong relationship with mentors dropouts

Predictive Analysis - Model

Algorithms used:

- Unsupervised learning (PCA and clustering): not intuitive and lower accuracy
- ✓ • Decision Tree: easier to understand and better accuracy

Note: some bootstrapping techniques were used to balance out the number of dropouts and formal closure in the dataset

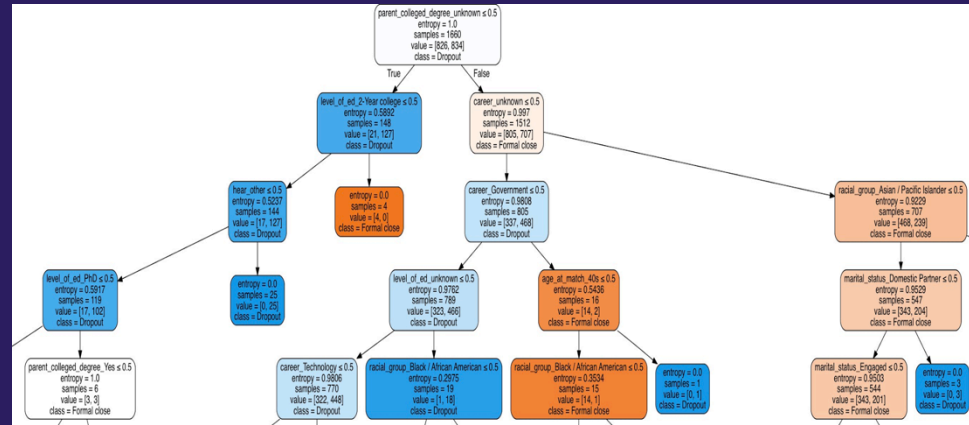


Figure 3 Extract of Decision Tree

Predictive analysis - Findings

Top 10 most important predictive indicators:

1. Career - not specified
2. Parent college degree - not specified
3. Heard iMentor from employer
4. Level of education - 2-yr college
5. Career in Government
6. Racial group - African American
7. Racial group - Asian
8. Career in Tech
9. Level of education - not specified
10. Age at match - 30's

Model accuracy: 65%

The low accuracy possible explanation:

- Demographic attributes of mentors alone can only partially explain mentors likelihood of dropping out. As next step, features other than demographics will need to be explored and added to the current model
- Small number of formal closure (i.e. the opposite label of mentor dropout) in the dataset in the 4 yr-program can weaken the predictive power of the model

Predictive analysis - Possible Uses

- Predict dropout probability of current mentors and detect potential group likely to dropout
 - Can be visualized on a map by school - see Figure B - Tableau demo (note mentors mailing zip code was used since school zip codes were not available)
- Better understand factors related to mentors dropout to help better mentor recruitment
- The prediction of mentors dropouts can also help for mentors population projection and resource planning

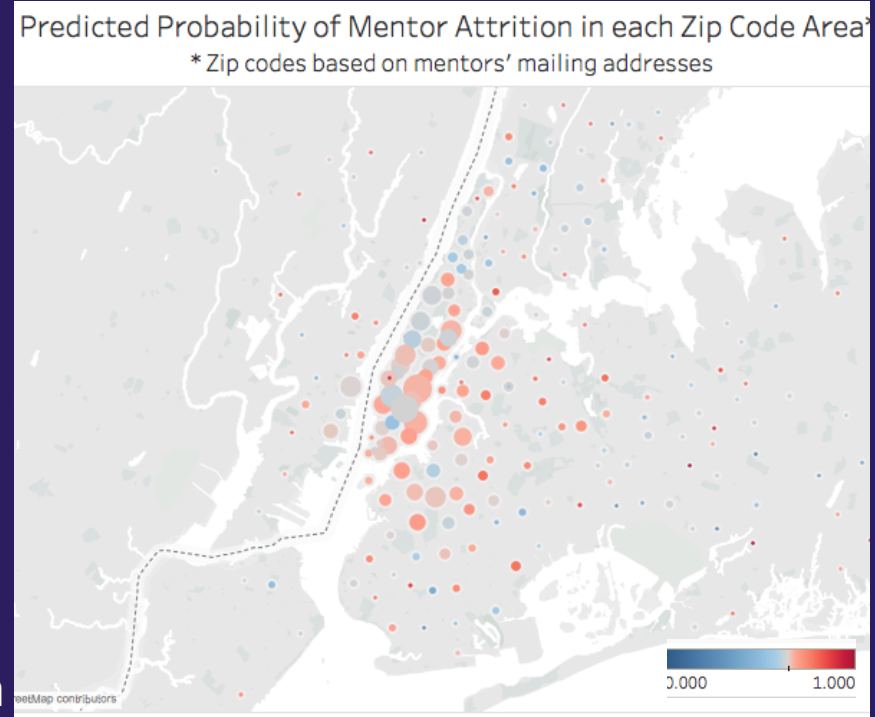


Figure B Map of mentors likelihood to drop out

4. Message traffic data Time Series Analysis

Key Findings

- No readily apparent behaviors distinguishing dropout mentors from other groups
- Patterns suggesting decreasing mentor engagement over time

Message Traffic Data Description

- Identifier for what week / lesson sequence each message was for
- Time stamp for login and send
- Message length
- Mentor, mentee, and pair identifiers
- School year 2015-2016
- Dataset for mentee and mentor messages, both with about 90,000 observations



Mentor Groups

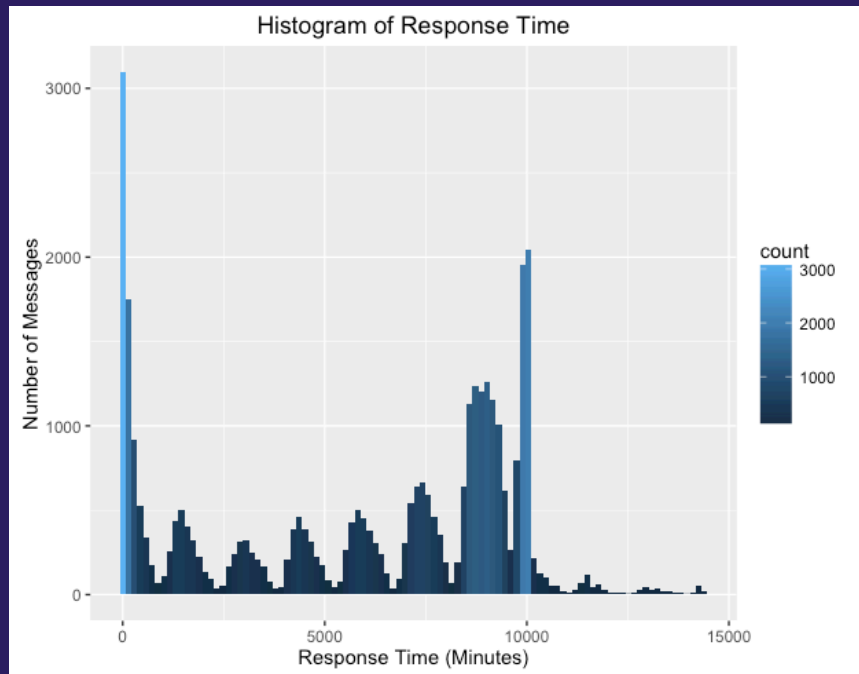
- Mentor dropout
- Formal closures (program completion)
- Mentee dropout
- Open matches

Methodology

- Compare metrics across groups
- Standardize measures calculating z score for each observation relative to mean and standard deviation for that particular mentor
- Group by message sequence to observe trends leading up to match closure or end of observations

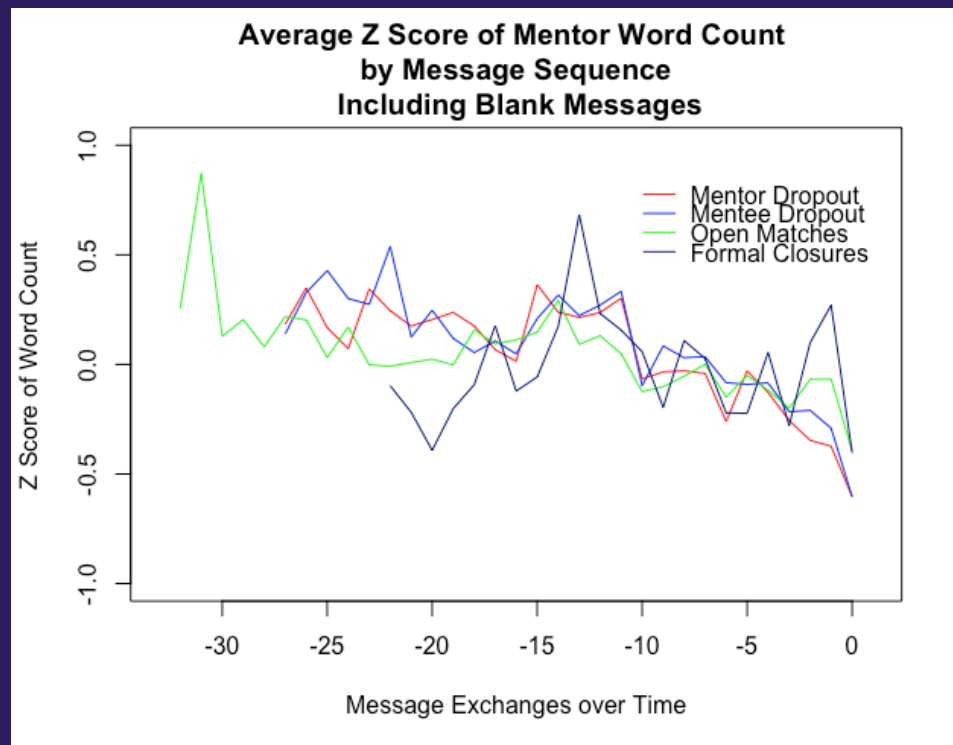
Response Time

- Large number of mentors who respond quickly each week
- Large spike as mentors rush to respond at the end of each week
- Hoped to see more last minute responses for dropout mentors, but the distribution looked similar for all mentor groups



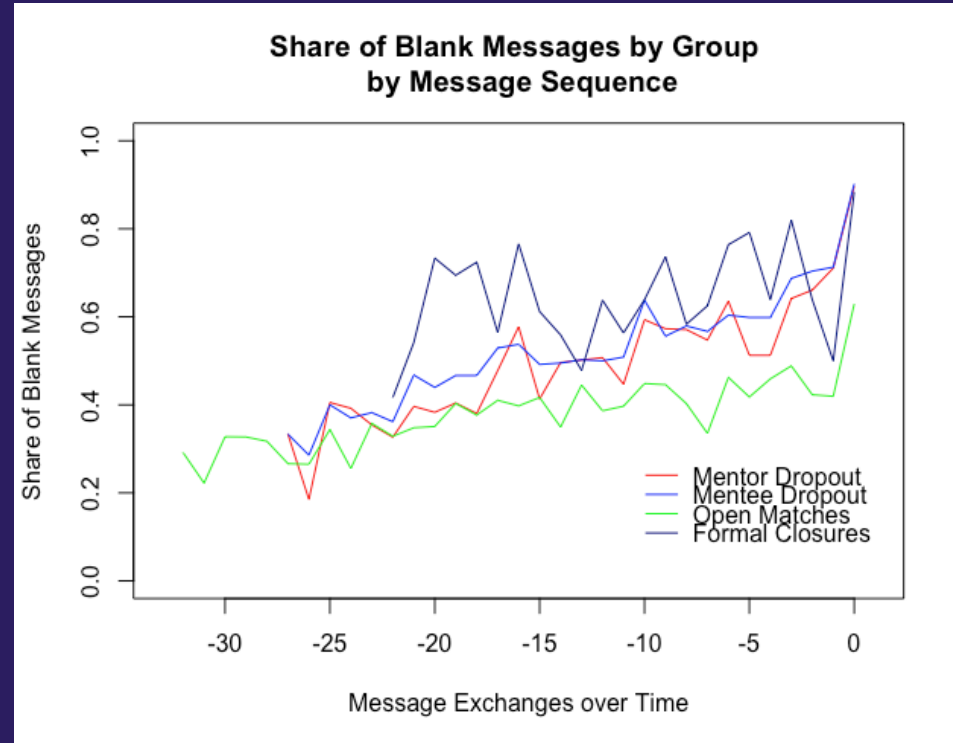
Word Count

- Mentors write on average about two and a half times as many words as mentees, with mean message length for mentors of 191 words and 74 words for mentees.
- Word count trends downward as mentors spend a longer time in the program



Blank Messages

- Very distinct increase in blank messages over time for all groups
- At least 40% of messages were blank for all groups
- Strangely, formal closures had highest prevalence of blank messages, difference was statistically significant at the $p < .001$ level using pairwise t-test with Bonferroni correction



Conclusion

- No strong indicators found to differentiate dropouts from formal closure:
 - Demographic attributes can partially train a prediction model but additional features should be added to complete the model
 - Message traffic analysis is based on one school year. Analysis over several years may reveal distinctions between dropouts and mentors who reach completion of the program
-

Proposed Next Steps

Data Architecture & Processing

- Will need to work with raw files that are used to generate the master data file.
- Need ability to “customize” tables for analysis -- data is currently heavily filtered manually
- Message logs and content will increase rapidly over time. May need to investigate Hadoop for stability
- Push demographics analysis back to cloud, create/update mentor dropout table with Bayesian statistics

Proposed Next Steps

Demographics Analysis

- Combine demographic attributes of mentors with features that reflect mentors behavior within program (e.g. message frequencies), mentors interaction and relationship with mentees (same interests, hobbies, family background, racial group, languages, etc.) and other features (proximity of work/home to school, etc.)

Proposed Next Steps

Messages & Logs

- Analyze patterns in chat, could be some interaction with blank messages
- If changes are made to the program, repeat analysis to see if they seem to improve mentor engagement
- Find out whether mentors begin to feel some fatigue over time
 - How do we explain consistent mentor disengagement?

The Team



Nicholas Chen

Time Series Analysis Lead

Message traffic analysis



Stephanie Fan

Data Visualization Lead

Tableau expert



Hyera Moon

Predictive Analysis Lead

Mentor demographic data
analysis



Leslie Teo

ETL Lead

Data architecture, storage
and retrieval expert