Mining the Common App

Mike Yung

in partnership with AdmitSee

mikeyung

yungmsh

yungmsh

Context

Part I: The Model

- Can we build a model that predicts a student's chances* of being admitted into college?

Part II: The Essay

- What insights can we glean from the Common App essay?

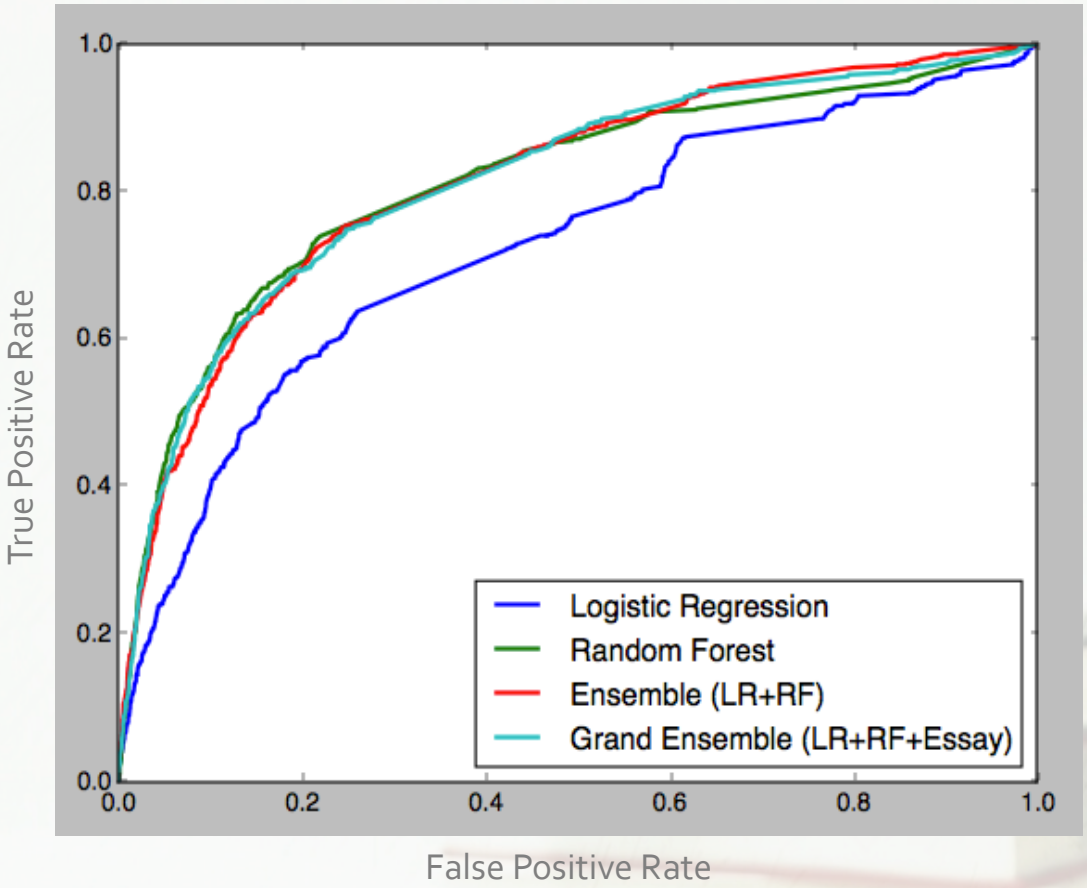
*There are some resources that 'calculate' your chances based on your GPA, SAT, and demographics, but none (at least publicly available) that take into account detailed factors such as specific extracurriculars, academic trajectory, the Common App essay etc.

Part 1: The Model

Can we build a model that predicts a student's chances of being admitted into college?

Evaluating the Model

ROC Curve (Test)



Test Set

Target Metric: ***Precision***

	Grand Ensemble	Ensemble	Random Forest
Accuracy	**88.1**	87.8	88.0
Precision	**62.8**	61.9	57.7
Recall	25.4	21.6	**37.1**

Interpreting the Model

Logistic Regression Model

Variable	$e^{Coefficient}$
Leader	2.26
Student Gov	1.69
Varsity Sport	1.58
Sports Captain	1.29
Award	1.22
Community Service	1.21
SAT Score	1.0003
SAT Times Taken	0.39

**Note: this is only a subset of all variables used in the model*

How to Interpret?

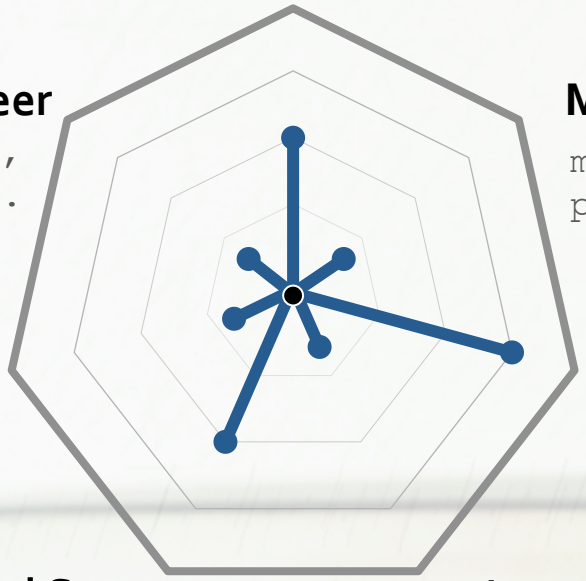
If you aren't already in a **leadership** position, taking one will **more than double** your odds of being admitted

Part 2: The Essay

What insights can we glean from the Common App essay?

What's in an Essay?

Topic Distribution of a Sample Essay



Family
mother, father, family, parent, sister, ...

Music/Arts
music, play, piano, perform, theater, ...

Sports
team, game, coach, player, season, ...

Culture
culture, world, language, travel, chinese, ...

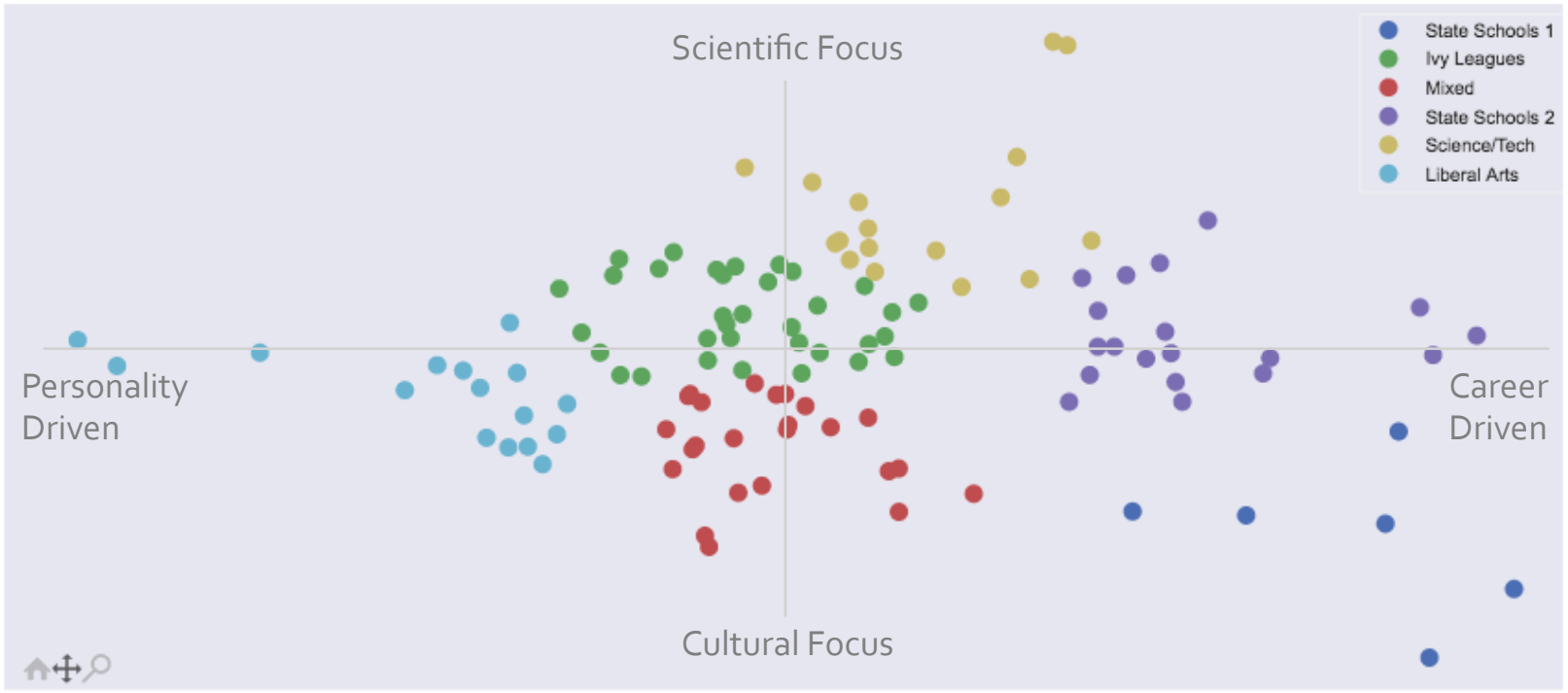
Personal Story
feel, think, friend, love, life, moment, ...

Science
research, science, technology, math, ...

Career
work, success, career, educate, community, ...

A 2-D Representation of College Essays

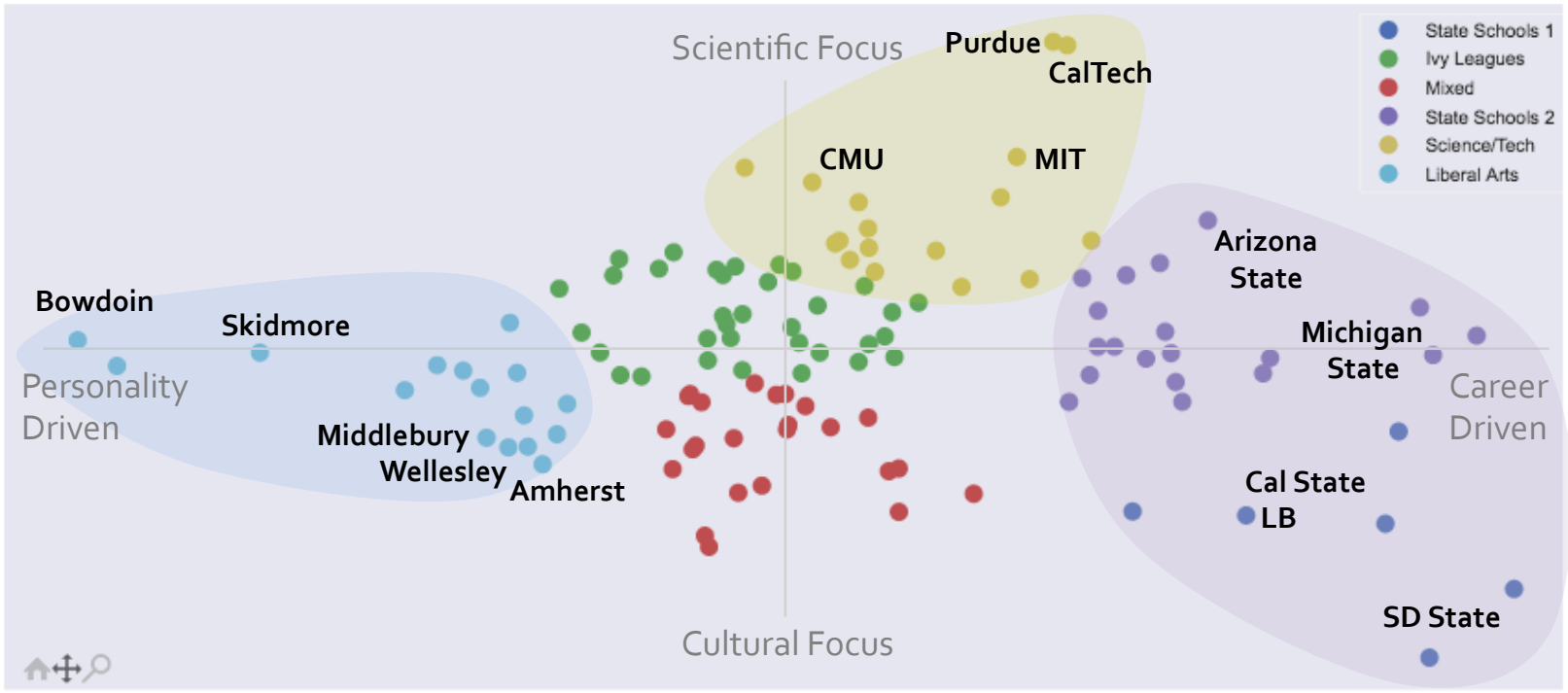
Step	Vectorize Text	Topic Modeling	**Visualize in 2D**	**Clustering**
Method	*TF-IDF*	*NMF*	***PCA***	***K-Means***



*Each point represents a college/university

A 2-D Representation of College Essays

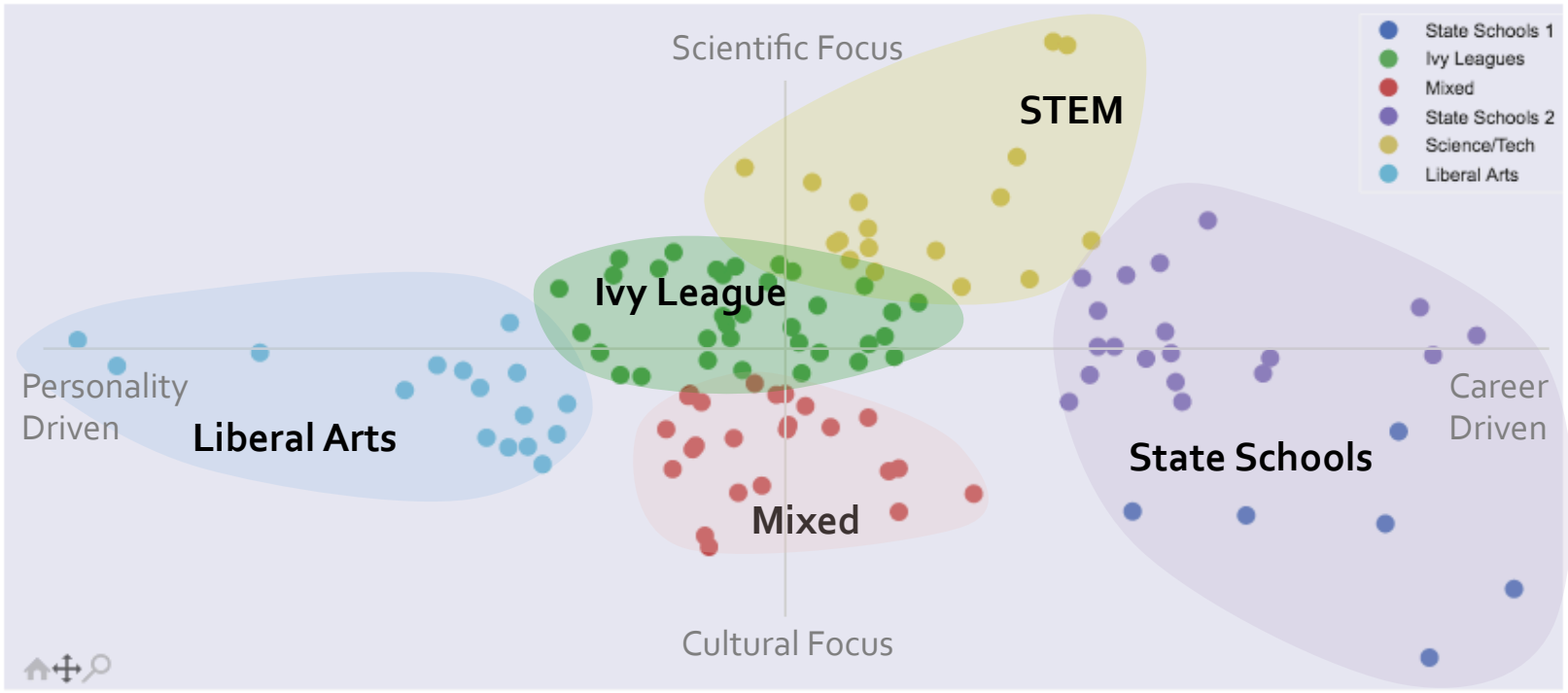
Step	Vectorize Text	Topic Modeling	**Visualize in 2D**	**Clustering**
Method	*TF-IDF*	*NMF*	***PCA***	***K-Means***



*Each point represents a college/university

A 2-D Representation of College Essays

Step	Vectorize Text	Topic Modeling	**Visualize in 2D**	**Clustering**
Method	*TF-IDF*	*NMF*	***PCA***	***K-Means***



*Each point represents a college/university

Final Thoughts

- Limitation of data
  - Only enough to model 'top school' admittance
  - With more data:
    - School-level model
    - Graduate school model
- Explore deeper feature-engineering
  - Interaction effects (e.g. Varsity*Captain)
  - Deeper effects (e.g. Hispanic student leading an African-American society)
- Refine topic modeling with LDA

Did You Know?

- If you gain an **additional 100 points** on your **SAT,** you can **increase** your odds of being admitted by **3%**
- If you **take the SATs one more time,** you can **reduce** your odds of being admitted by **61%**

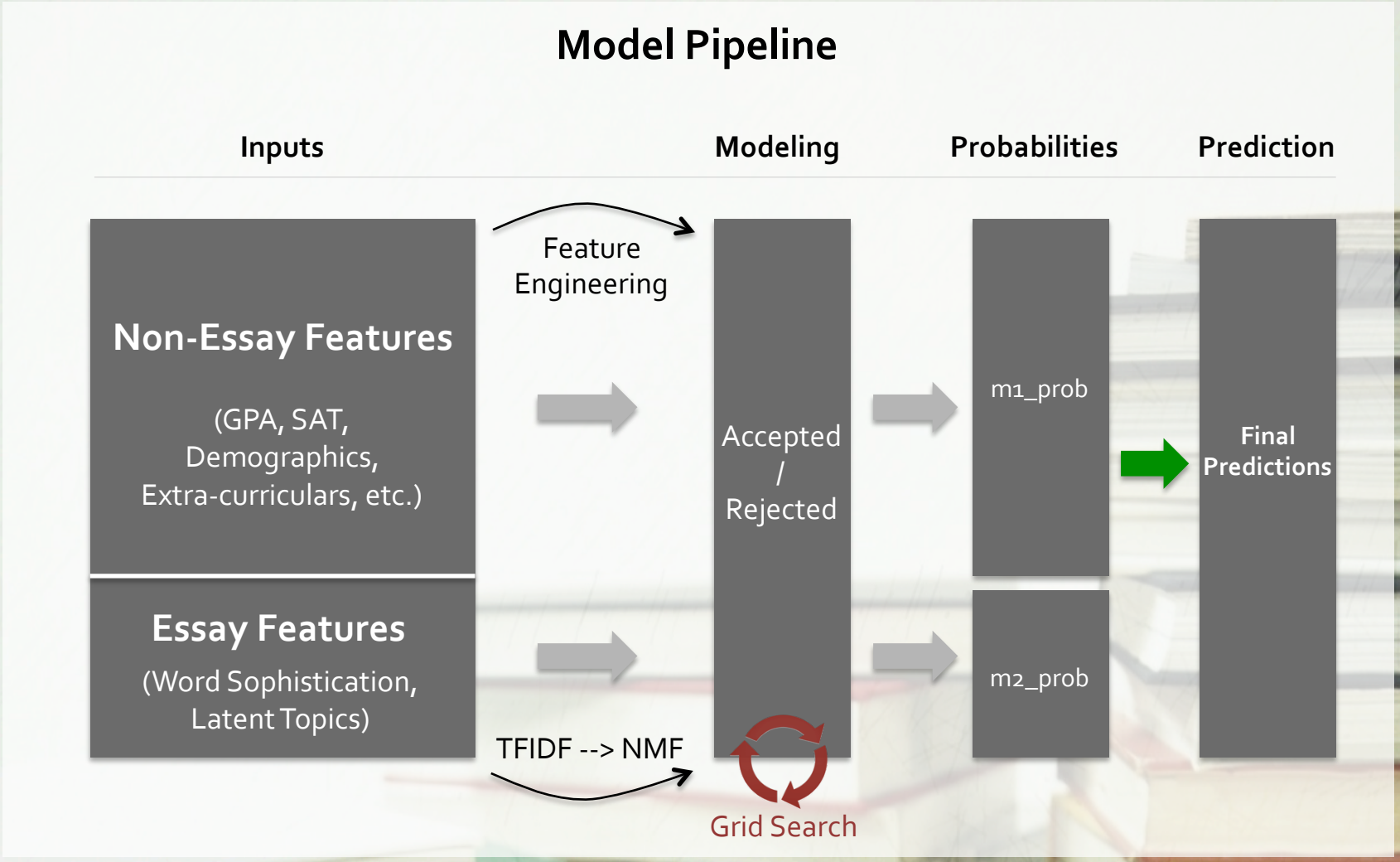
Thank You

mikeyung

yungmsh

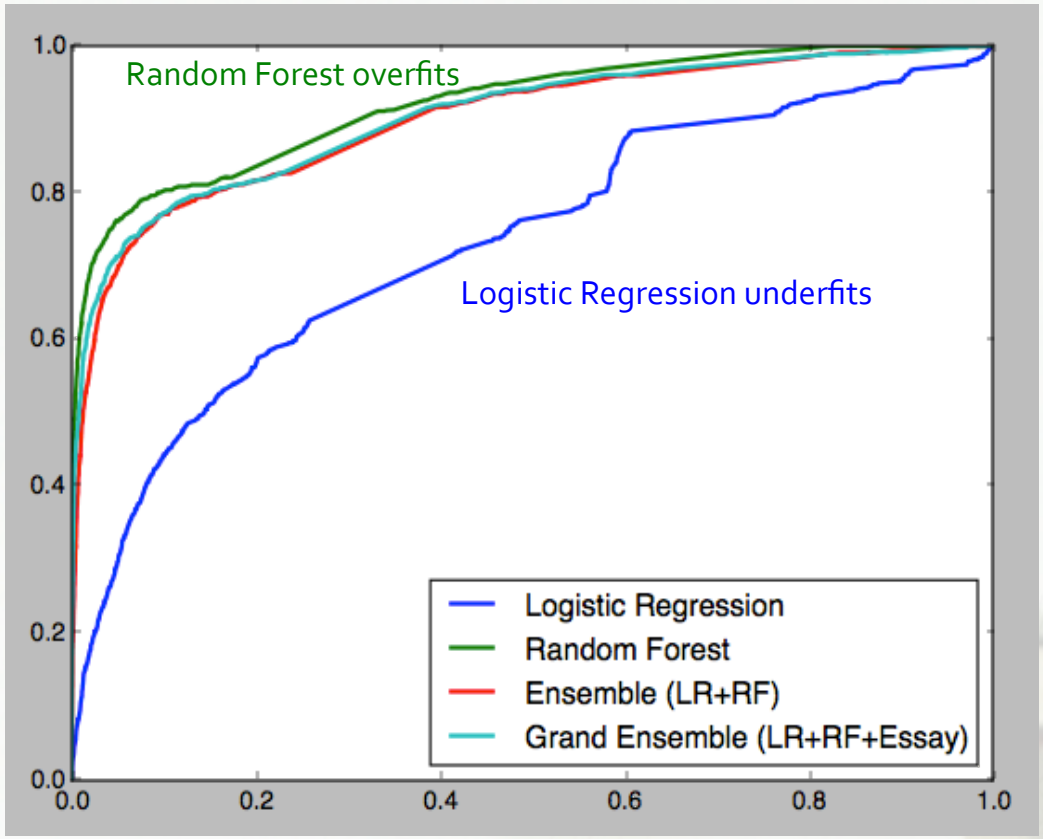
yungmsh

Appendix 1: Ensemble Model Pipeline



Appendix 2: ROC Curve for Train Set

ROC Curve (Train)



Train Set

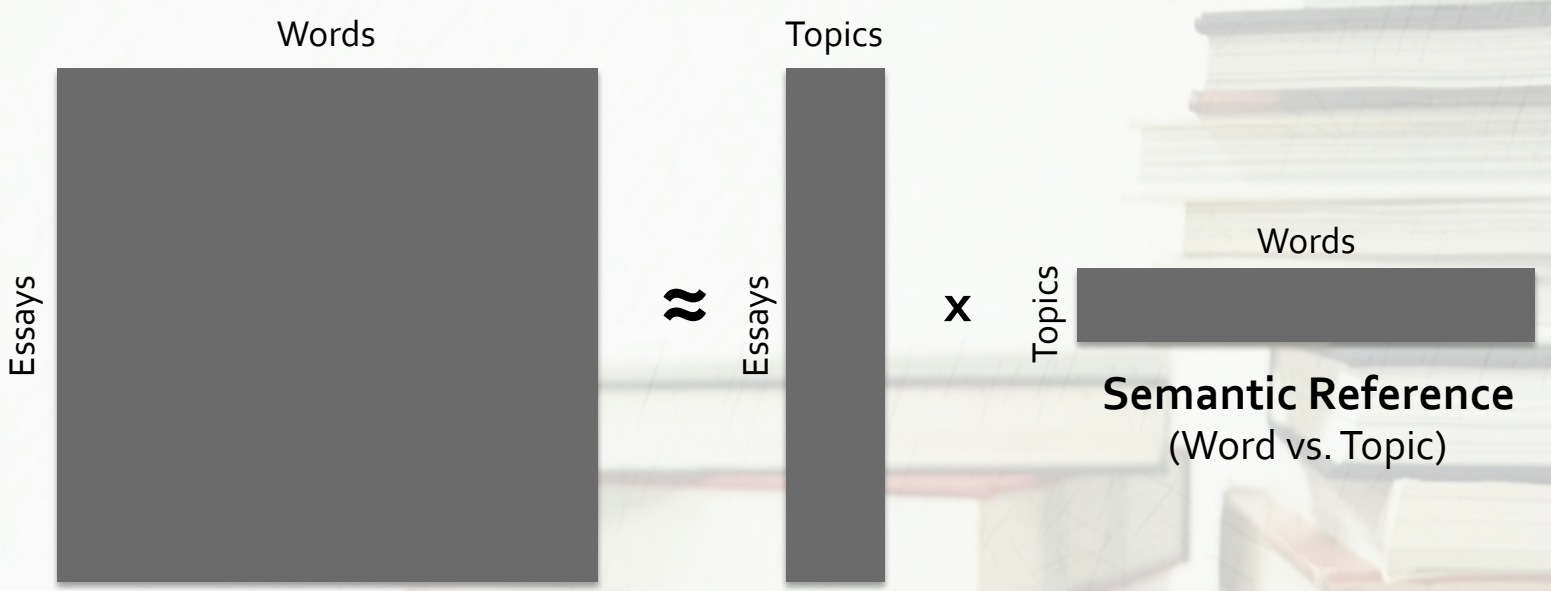
Target Metric: ***Precision***

	Grand Ensemble	Ensemble	Random Forest
Accuracy	92.4	91.2	**94.2**
Precision	**93.1**	91.2	90.1
Recall	45.7	37.2	**63.3**

Appendix 3: NMF Visualized

Step	**Vectorize Text**	**Topic Modeling**	Analyze
Method	***TF-IDF***	***NMF (Fit)***	*NMF (Transform)*

Visualization of NMF



Old Matrix
(Word vs. Essay)

New Matrix
(Topic vs. Essay)

Semantic Reference
(Word vs. Topic)

Appendix 4: NMF Semantic Reference Table

Step	**Vectorize Text**	**Topic Modeling**	Analyze
Method	***TF-IDF***	***NMF (Fit)***	*NMF (Transform)*

Non-Negative Matrix Factorization

Words

Topics							
mother	father	family	parent	sister	....	**Family**
music	play	piano	perform	theater	...	**Music/Arts**
culture	world	language	travel	american	...	**Culture**
team	game	coach	player	season	...	**Sports**
feel	think	friend	love	moment	...	**Personal/Story**
research	science	computer	technology	math	...	**Science**
work	education	career	success	community	...	**Career**

Semantic Reference
(Word vs. Topic)