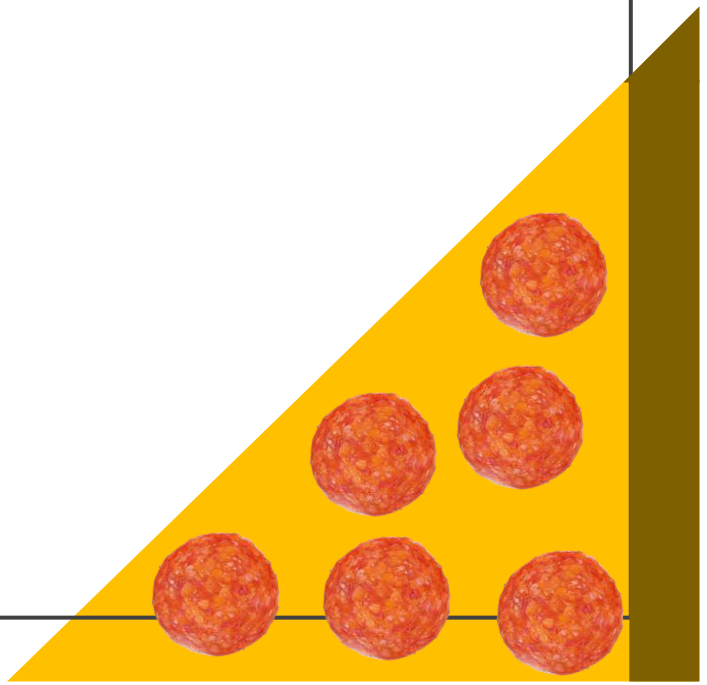
RAOP - Restoring Faith in Humanity, One Slice at a Time

r/Random_Acts_Of_Pizza

Join

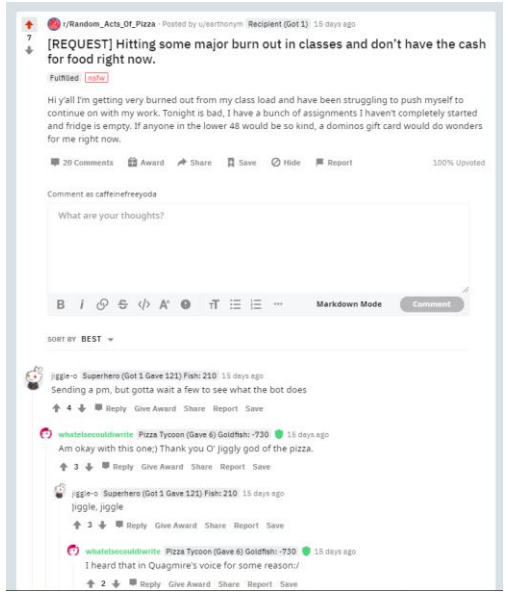
Predicting Pizzas

Amber Chen, Dustin Cox, Heather Heck, & Laura Treider



Recipe

- 1 Kaggle Challenge
- 1 Team Approach
- 4 Parts Exploratory Data Analysis
- 1 Dash of Feature Engineering
- 5 Rounds of Modeling
- 1 Part Performance Summary
- 1 Heap of Learning
- Bake in it for a few weeks, and enjoy!

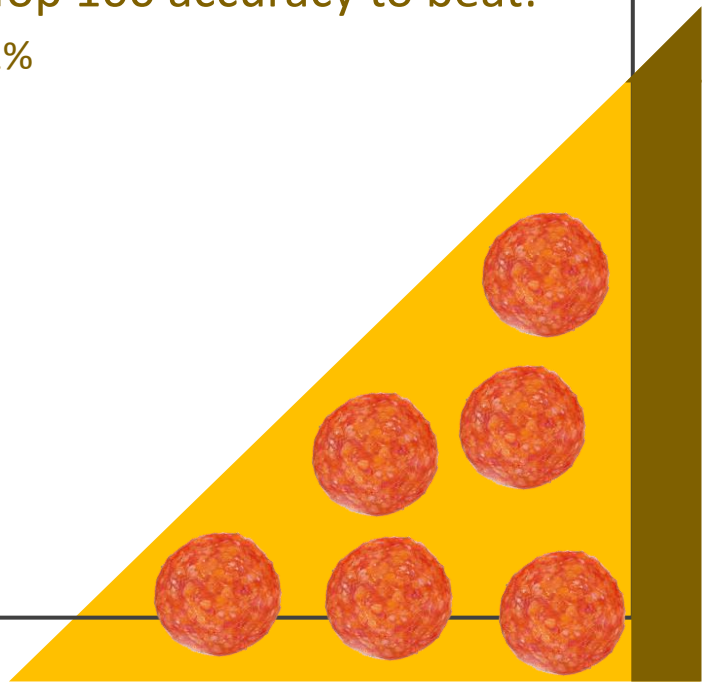


```
raw train_data example:

{'giver_username_if_known': 'N/A',
 'number_of_downvotes_of_request_at_retrieval': 0,
 'number_of_upvotes_of_request_at_retrieval': 1,
 'post_was_edited': False,
 'request_id': 't3_125d7',
 'request_number_of_comments_at_retrieval': 0,
 'request_text': 'Hi I am in need of food for my 4 children we are a military '
                 'family that has really hit hard times and we have exahusted '
                 'all means of help just to be able to feed my family and make '
                 'it through another night is all i ask i know our blessing is '
                 'coming so whatever u can find in your heart to give is '
                 'greatly appreciated',
 'request_text_edit_aware': 'Hi I am in need of food for my 4 children we are '
                            'a military family that has really hit hard times '
                            'and we have exahusted all means of help just to '
                            'be able to feed my family and make it through '
                            'another night is all i ask i know our blessing is '
                            'coming so whatever u can find in your heart to '
                            'give is greatly appreciated',
 'request_title': 'Request Colorado Springs Help Us Please',
 'requester_account_age_in_days_at_request': 0.0,
 'requester_account_age_in_days_at_retrieval': 792.4204050925925,
 'requester_days_since_first_post_on_raop_at_request': 0.0,
 'requester_days_since_first_post_on_raop_at_retrieval': 792.4204050925925,
 'requester_number_of_comments_at_request': 0,
 'requester_number_of_comments_at_retrieval': 0,
 'requester_number_of_comments_in_raop_at_request': 0,
 'requester_number_of_comments_in_raop_at_retrieval': 0,
 'requester_number_of_posts_at_request': 0,
 'requester_number_of_posts_at_retrieval': 1,
 'requester_number_of_posts_on_raop_at_request': 0,
 'requester_number_of_posts_on_raop_at_retrieval': 1,
 'requester_number_of_subreddits_at_request': 0,
 'requester_received_pizza': False,
 'requester_subreddits_at_request': [],
 'requester_upvotes_minus_downvotes_at_request': 0,
 'requester_upvotes_minus_downvotes_at_retrieval': 1,
 'requester_upvotes_plus_downvotes_at_request': 0,
 'requester_upvotes_plus_downvotes_at_retrieval': 1,
 'requester_user_flair': None,
 'requester_username': 'nickylvst',
 'unix_timestamp_of_request': 1317852607.0,
 'unix_timestamp_of_request_utc': 1317849007.0}
-----------------------------------------------
```

- 4,040 train data examples
- 1,631 test data examples
- 2 outcome categories to predict
  - Pizza!
  - No pizza. ☹
- Kaggle Top 100 accuracy to beat:
  - 69.2%

The Kaggle Challenge



Team Approach

Exploratory Data Analysis
- Understand fields
- Find data quality issues
- Uncover interesting relationships

Baseline Model
- Guess all 1 category to determine the underlying distribution and simplest model to beat

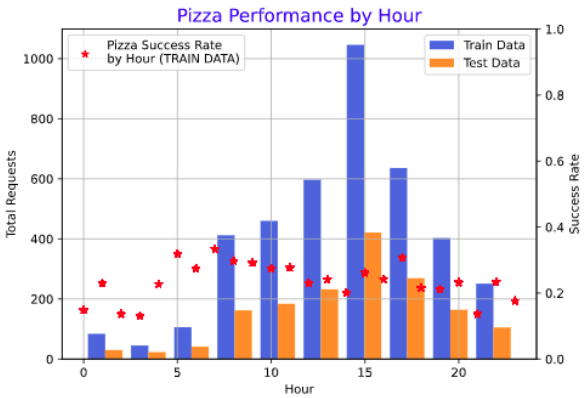
Engineer Features
- Decompose from existing features
- Pre-process fields
- Cluster analyses to group logically

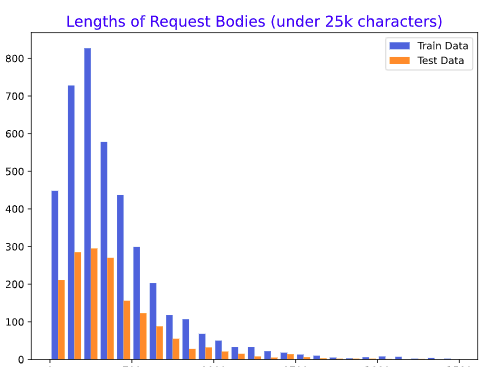
Iterate through various models
- Regression (linear / logistic)
- Decision tree / forest
- …

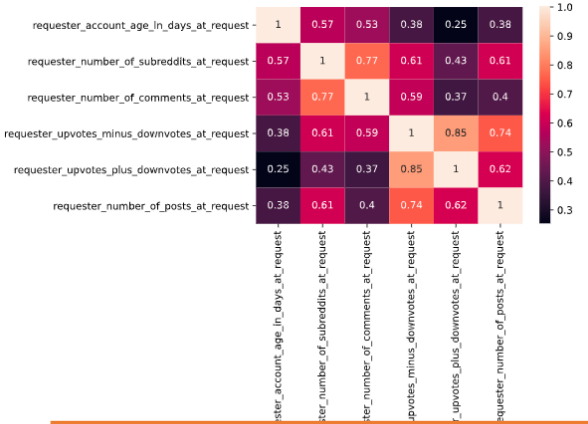
Iterate through Ensembles
- Combine promising and complimentary models
- Tune hyperparameters for test data accuracy

Select Best-Performing Model
- We have a winner!

Exploratory Data Analysis







Train vs. Test Data Differences

- Train: Everything Pizza
- Test: Pepperoni Pizza
- i.e., various columns in the train set not present in test, and therefore unusable

Pizza Success Rates

- Time of day
- Half of month

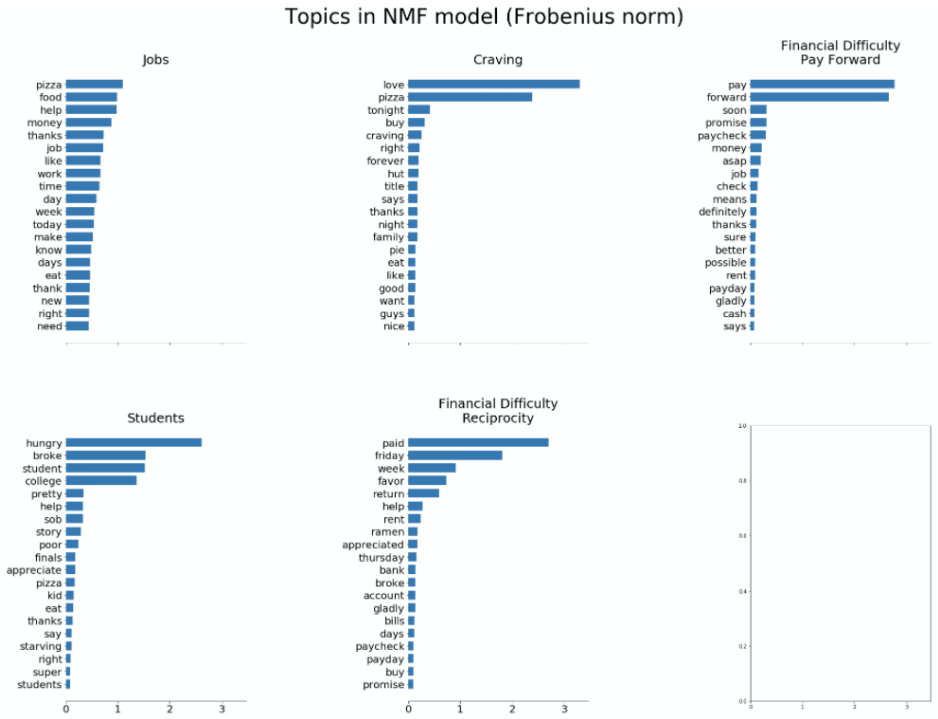
Post Submission Attributes

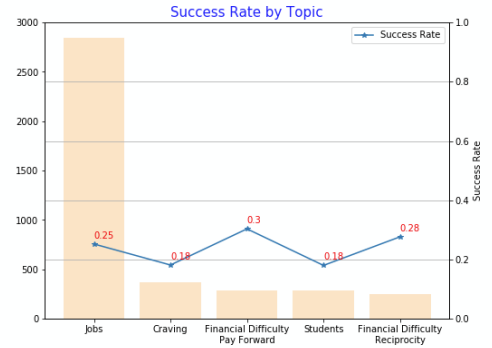
- Post lengths
- Contains image(s)
- Contains URL(s)
- Upvotes / downvotes

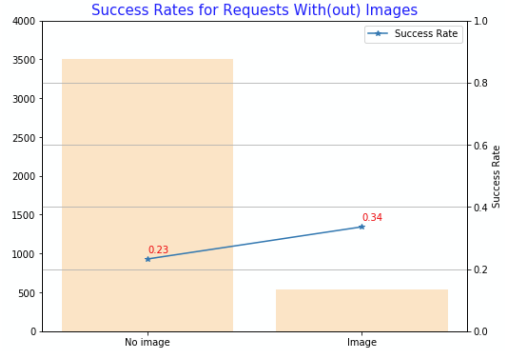
Poster Attributes

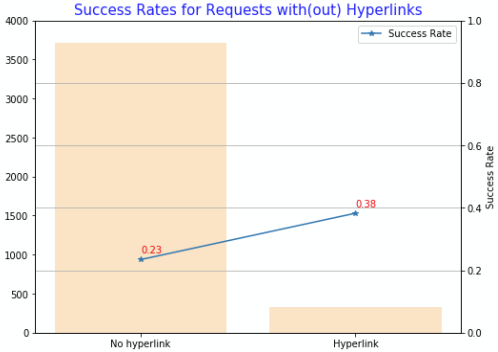
- Throwaway accounts
- Highly active users
  - Posts
  - Comments
  - Subreddit membership

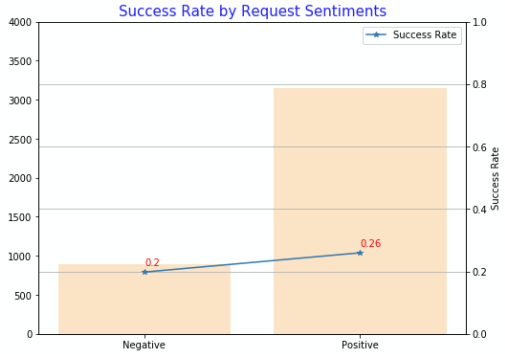
Feature Engineering









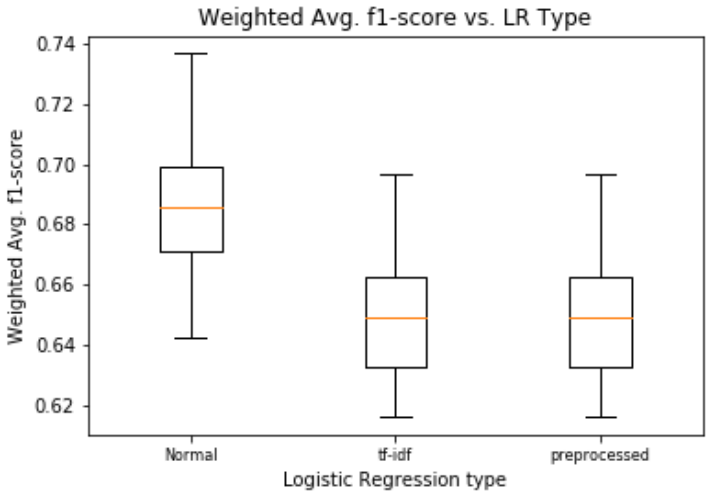


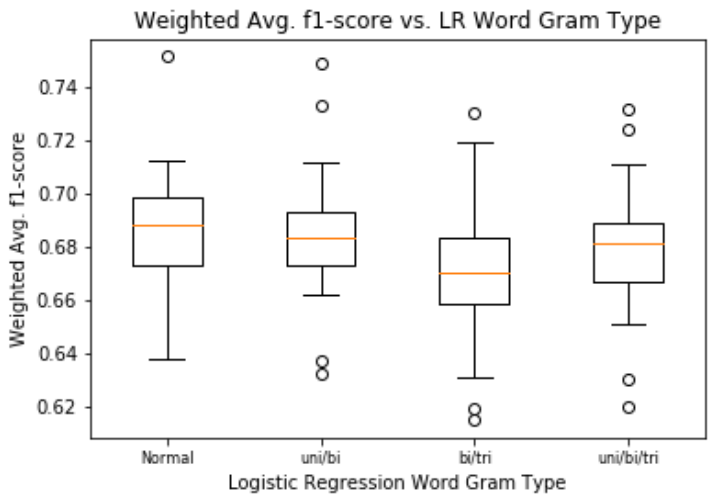
Textual Analysis on Requesters' Posts

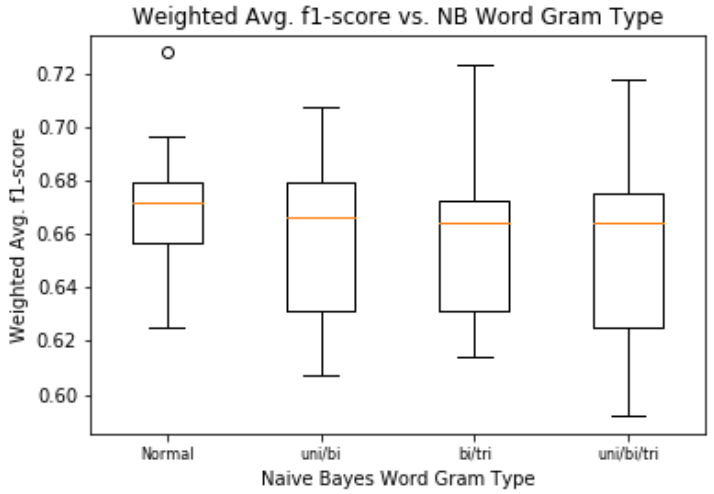
- Contains image(s)
- Contains URL(s)
- 5 topics defined using K-means, PCA and NMF models
- Sentiment analysis

Pizza Success Rates

- Vary among different groups
- Higher for posts that include image(s)/URL(s)/positive sentiment







Modeling:
LR Tuning
Gram Bag

Request Text Only: Tuning LR Model / Gram Bag

- Used 25-fold cross-validation to optimize model to dev data.
- Model based on plain linear regression performed better than tf-idf, and also better than preprocessed text model.
- Unigrams weren't outperformed by combinations with bigrams and trigrams.
- Logistic regression of text outperformed the Naïve Bayes model for all word gram types.

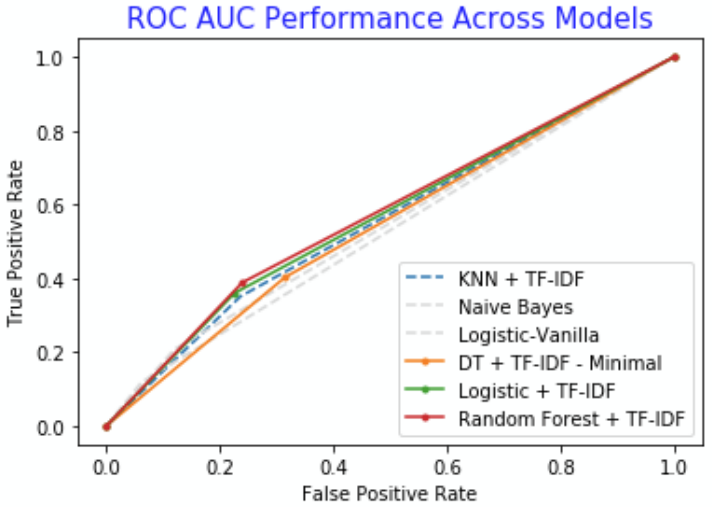
Modeling & Performance Summary

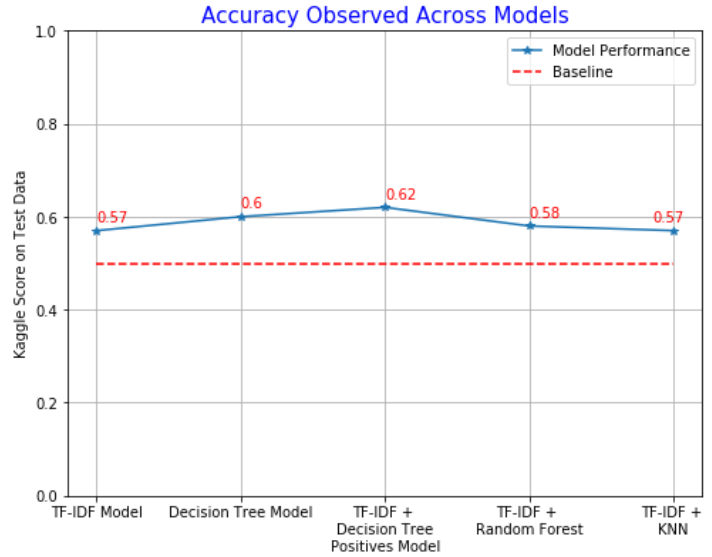
'requester_account_age_in_days_at_request',
'requester_days_since_first_post_on_raop_at_request',
'requester_number_of_comments_in_raop_at_request',
'requester_number_of_posts_on_raop_at_request',
'requester_upvotes_minus_downvotes_at_request',

Natural Features

'Length_of_Post',
'RAOP_member_at_time_of_request',
'zero_to_three',
'four_to_seven',
'eight_to_eleven',
'twelve_to_fifteen',
'sixteen_to_nineteen',
'twenty_to_twentythree',
'beg_month_of_request',
'includes_visuals',
'includes_hyperlink',
'Topic1_dec',
'Topic2_dec',
'Topic3_dec',
'Topic4_dec',
'Topic5_dec',
'sentiment_bin'

Engineered Features



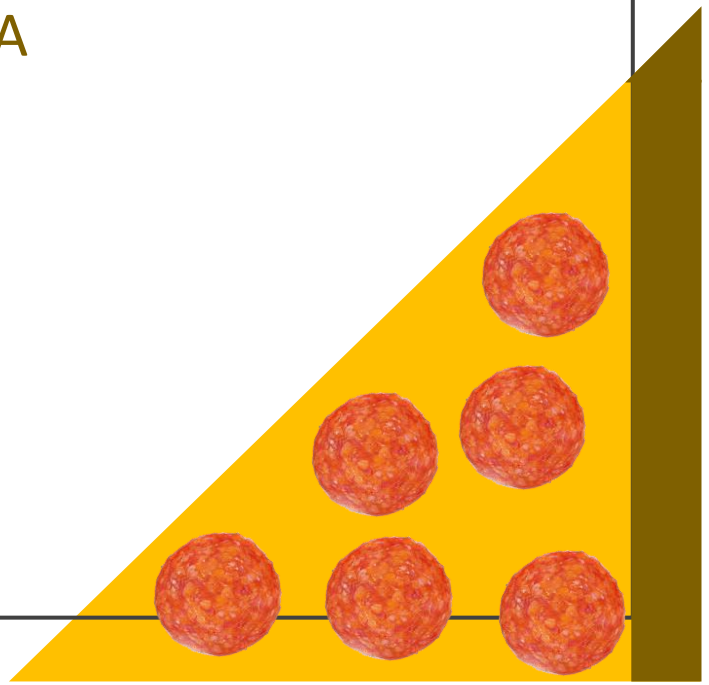


Learning

- Theory and Practice
  - Teaching grants access to understanding, range of tools and approaches
  - No substitute for practice and application
- Art and Science
  - Infinite combinations of features, models, hyperparameters, ensembles to try
  - What to go after is a combination of domain knowledge, experience, clues from the data, gut intuition, and creativity
- Train and Test
  - If train and test data look materially different, it opens a whole set of problems for which the team must correct

THANK YOU

Q&A



Team Approach

Exploratory Data Analysis

- Understand fields
- Find data quality issues
- Uncover interesting relationships

Baseline Model

- Guess all 1 category to determine the underlying distribution and simplest model to beat

Engineer Features

- Decompose from existing features
- Pre-process fields
- Cluster analyses to group logically

Iterate through various models

- Regression (linear / logistic)
- Decision tree / forest
- …

Iterate through Ensembles

- Combine promising and complimentary models
- Tune hyperparameters for test data accuracy

Select Best-Performing Model

- We have a winner!