Caterpillar Tube Pricing

Adam Cone and Ismael Cruz
Capstone Project
NYCDSA Bootcamp 005

Introduction

kaggle Host Competitions Datasets Scripts Jobs Community IsmaelCruz Logout

Completed • $30,000 • 1,323 teams

Caterpillar Tube Pricing

Mon 29 Jun 2015 – Mon 31 Aug 2015 (9 months ago)

Dashboard

Home
- Data
- Make a submission

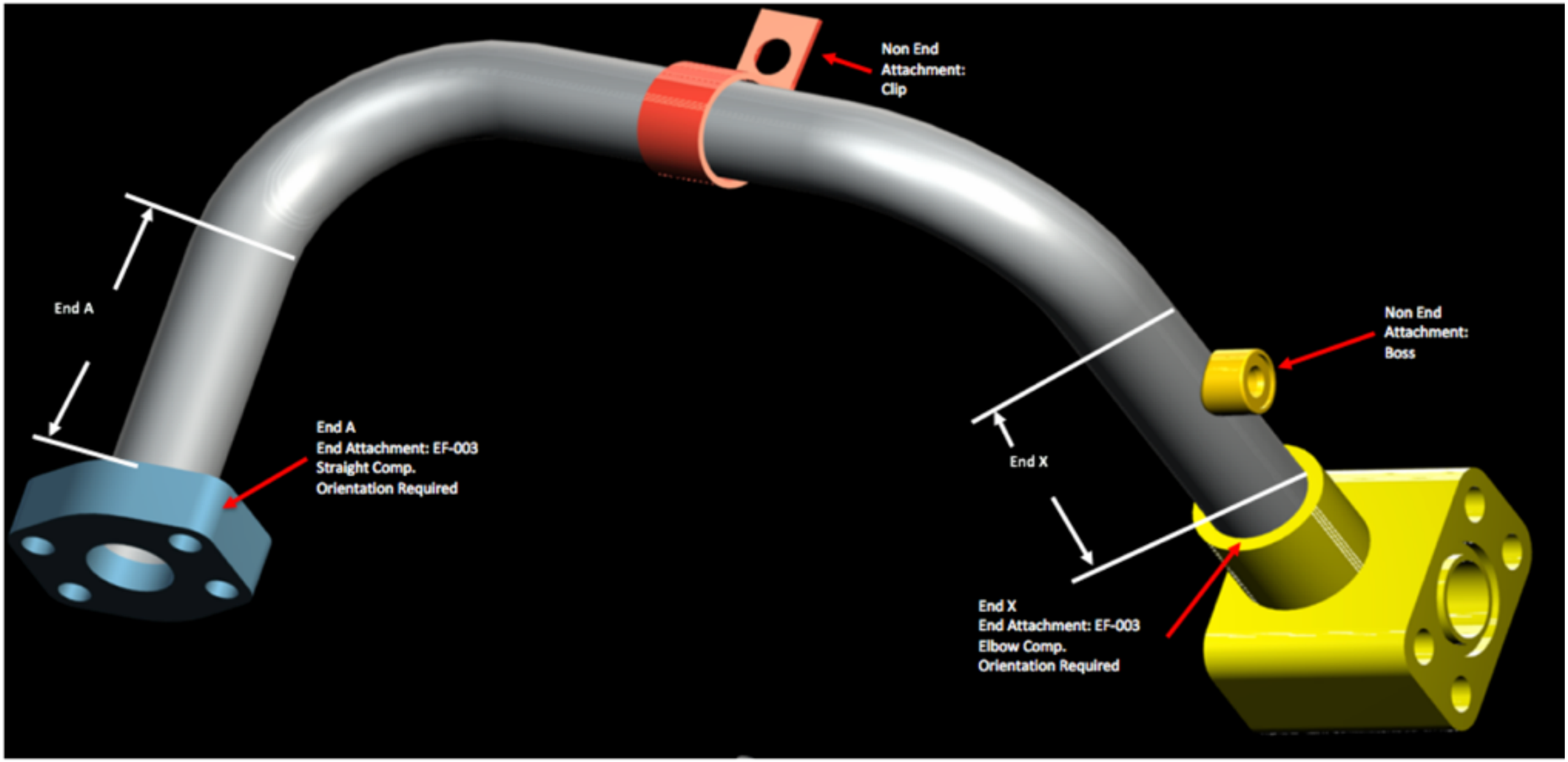
Information
- Description

Competition Details » Get the Data » Make a submission

Model quoted prices for industrial tube assemblies

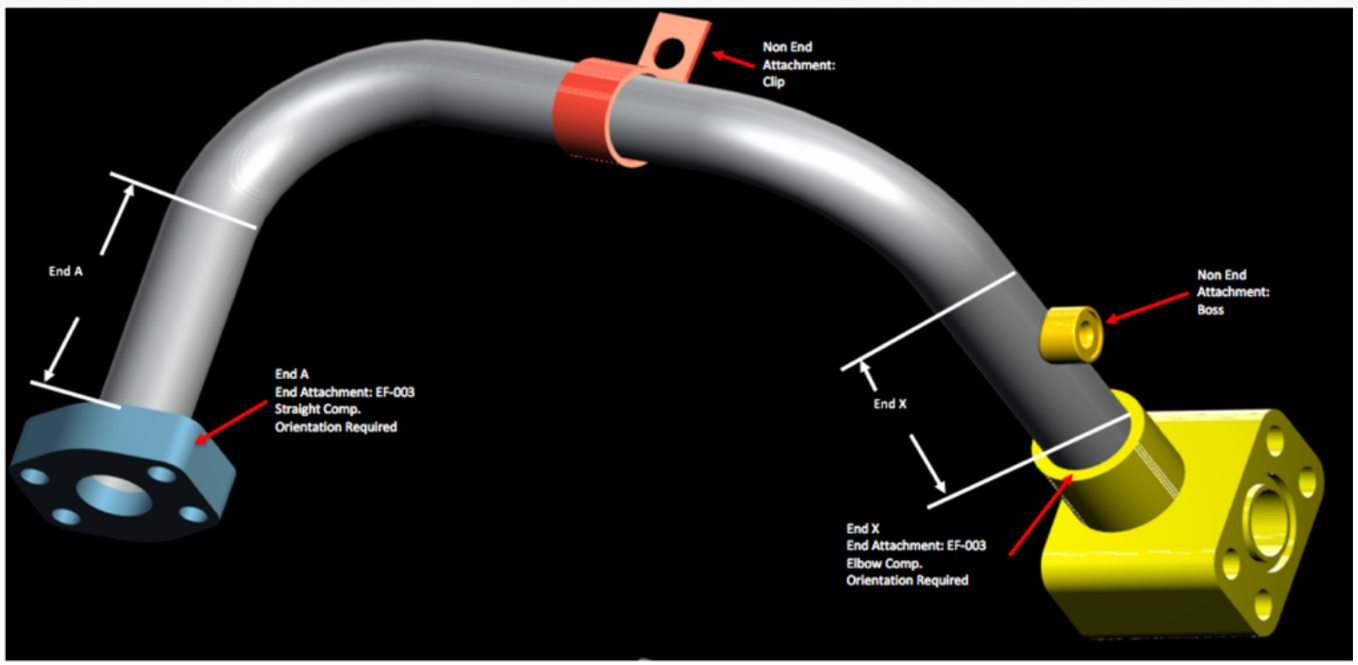
Introduction

Introduction



Data

```
bill_of_materials_df = pd.read_csv('../competition_data/bill_of_materials.csv')
comp_adaptor_df = pd.read_csv('../competition_data/comp_adaptor.csv')
comp_boss_df = pd.read_csv('../competition_data/comp_boss.csv')
comp_elbow_df = pd.read_csv('../competition_data/comp_elbow.csv')
comp_float_df = pd.read_csv('../competition_data/comp_float.csv')
comp_hfl_df = pd.read_csv('../competition_data/comp_hfl.csv')
comp_nut_df = pd.read_csv('../competition_data/comp_nut.csv')
comp_other_df = pd.read_csv('../competition_data/comp_other.csv')
comp_sleeve_df = pd.read_csv('../competition_data/comp_sleeve.csv')
comp_straight_df = pd.read_csv('../competition_data/comp_straight.csv')
comp_tee_df = pd.read_csv('../competition_data/comp_tee.csv')
comp_threaded_df = pd.read_csv('../competition_data/comp_threaded.csv')
components_df = pd.read_csv('../competition_data/components.csv')
specs_df = pd.read_csv('../competition_data/specs.csv')
test_set_df = pd.read_csv('../competition_data/test_set.csv')
train_set_df = pd.read_csv('../competition_data/train_set.csv')
tube_end_form_df = pd.read_csv('../competition_data/tube_end_form.csv')
tube_df = pd.read_csv('../competition_data/tube.csv')
type_component_df = pd.read_csv('../competition_data/type_component.csv')
type_connection_df = pd.read_csv('../competition_data/type_connection.csv')
type_end_form_df = pd.read_csv('../competition_data/type_end_form.csv')
```



Data

data issue	solution

Data

data issue	solution
NaNs in categorical data	new NaN factor level

Data

data issue	solution
NaNs in categorical data	new NaN factor level
NaNs in numerical data 30 of 60,000 in bend_radius	impute with mean

Data

data issue	solution
NaNs in categorical data	new NaN factor level
NaNs in numerical data 30 of 60,000 in bend_radius	impute with mean
factor variables can't be used for analysis	convert to dummy variables

Data

data issue	solution
NaNs in categorical data	new NaN factor level
NaNs in numerical data 30 of 60,000 in bend_radius	impute with mean
factor variables can't be used for analysis	convert to dummy variables
order date format can't be used for analysis	convert to days since earliest date

Data

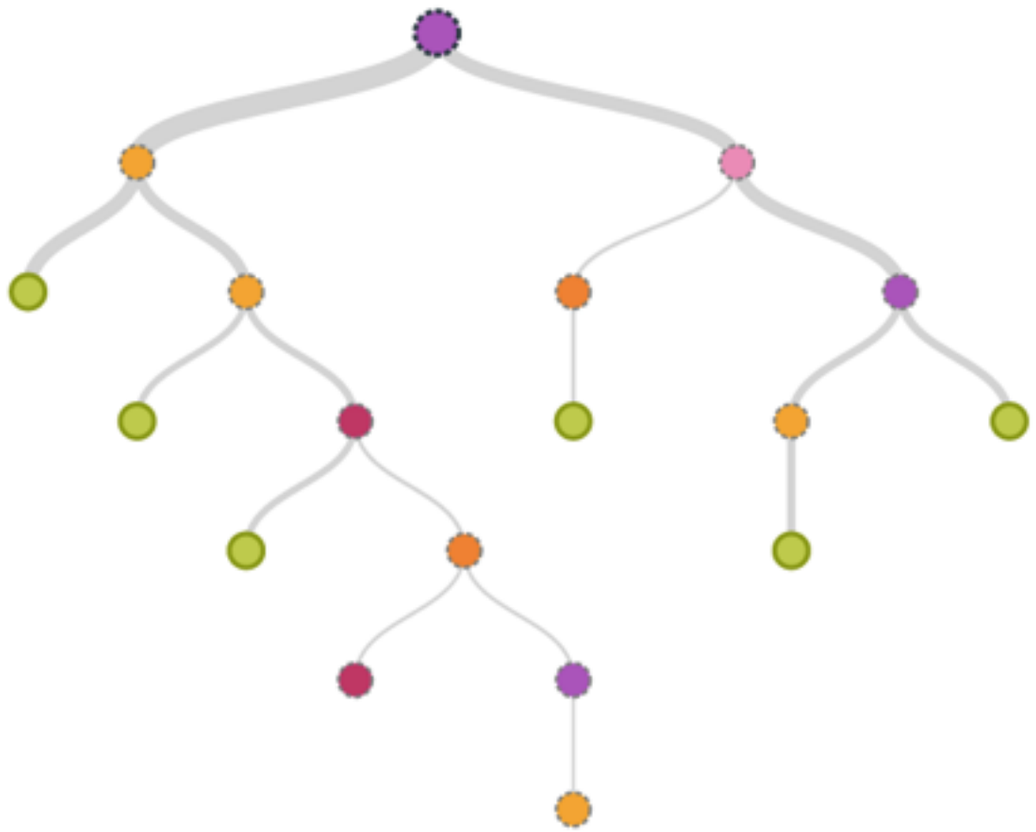
data issue	solution
NaNs in categorical data	new NaN factor level
NaNs in numerical data 30 of 60,000 in bend_radius	impute with mean
factor variables can't be used for analysis	convert to dummy variables
order date format can't be used for analysis	convert to days since earliest date
tube assemblies have different components	feature engineering (summary & detailed)

Data

training data frame features

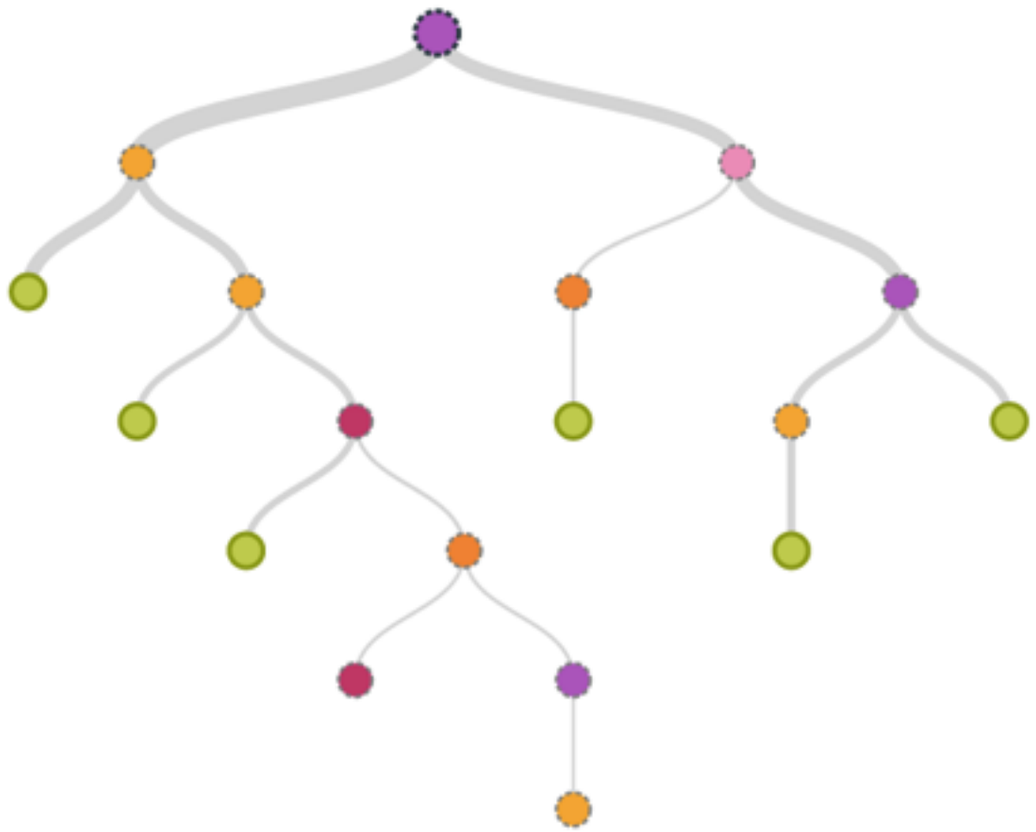
basic	components
158	169
basic FE	basic FE & detailed FE

Modeling



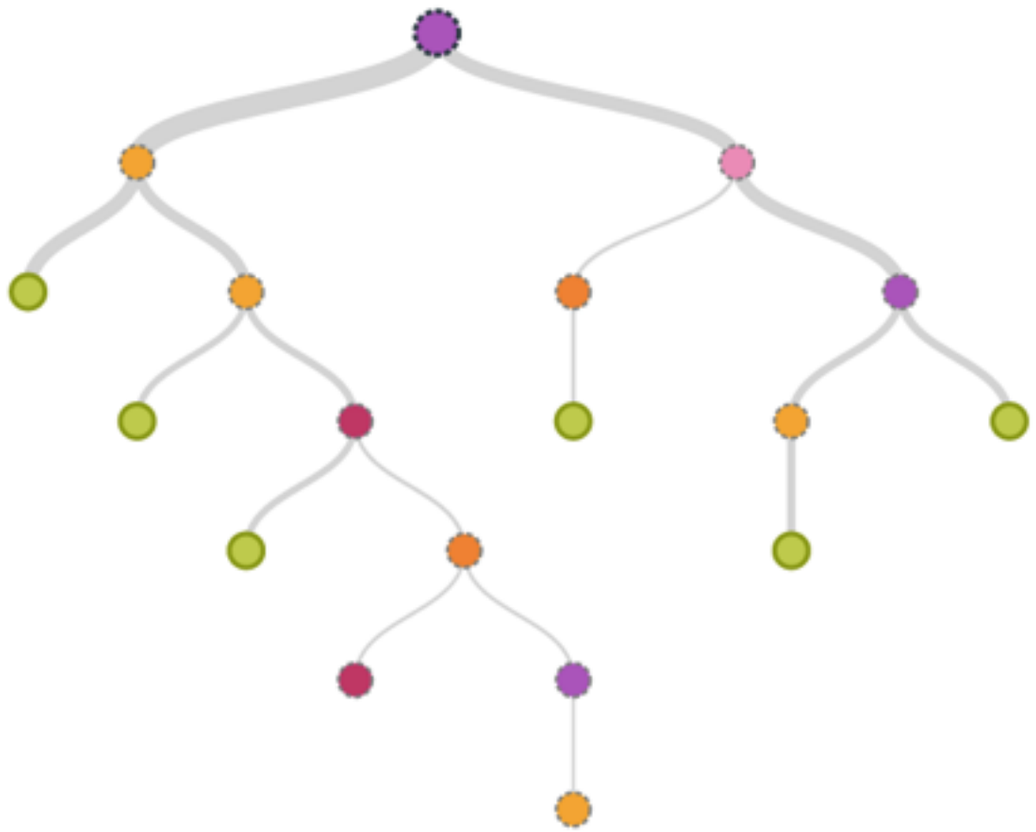
1. decision trees
2. random forest
3. gradient boosting

Modeling



Expectations

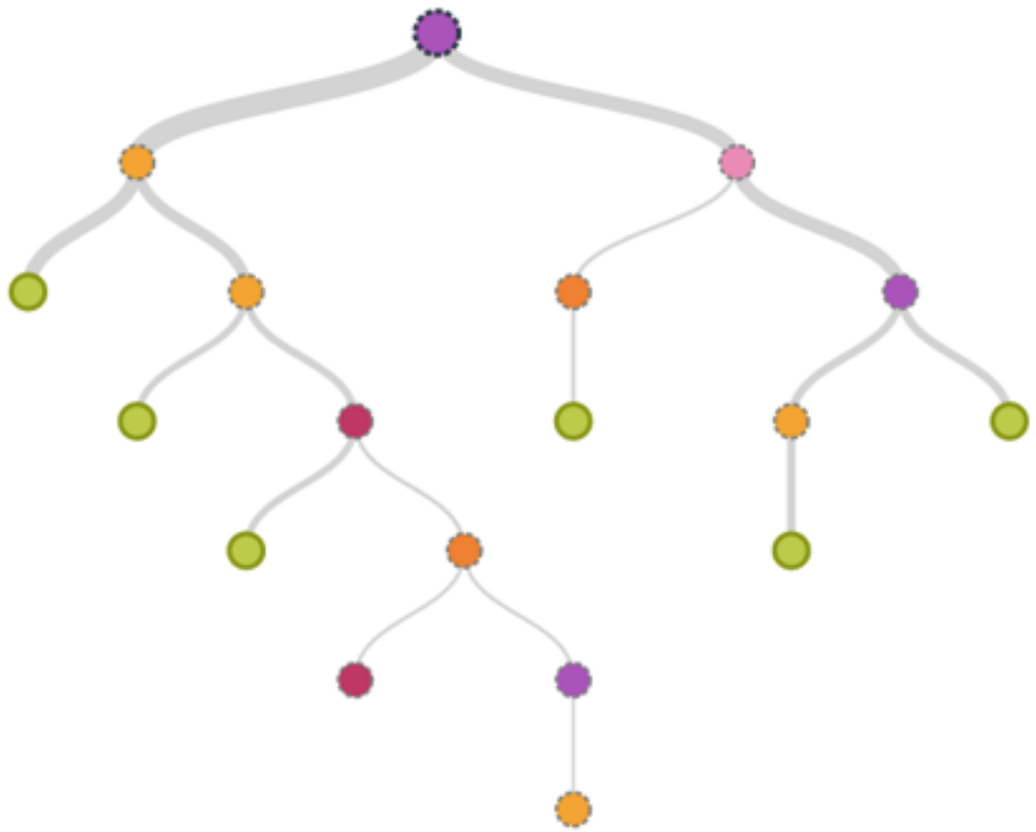
Modeling



<u>Expectations</u>

1. component > basic

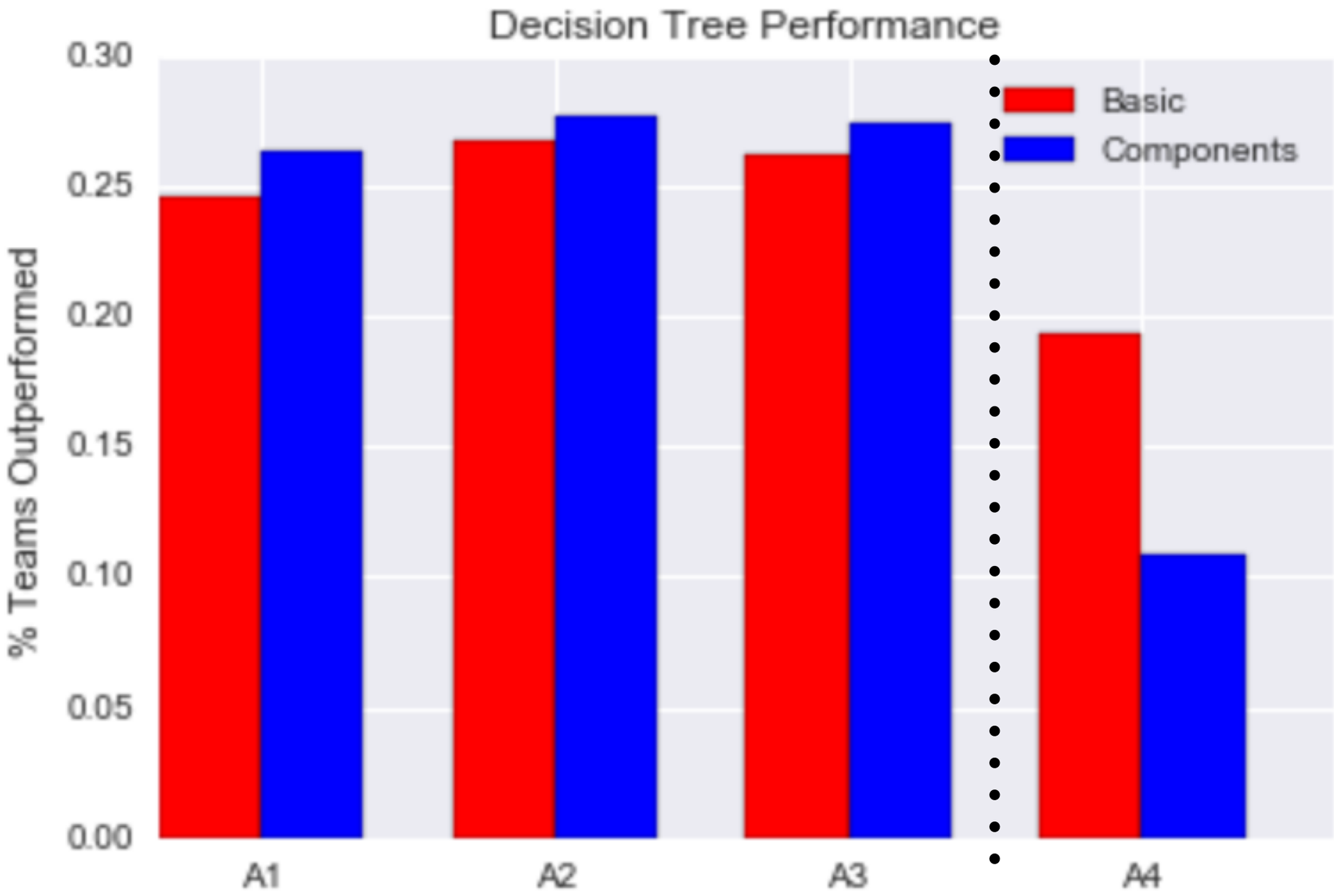
Modeling



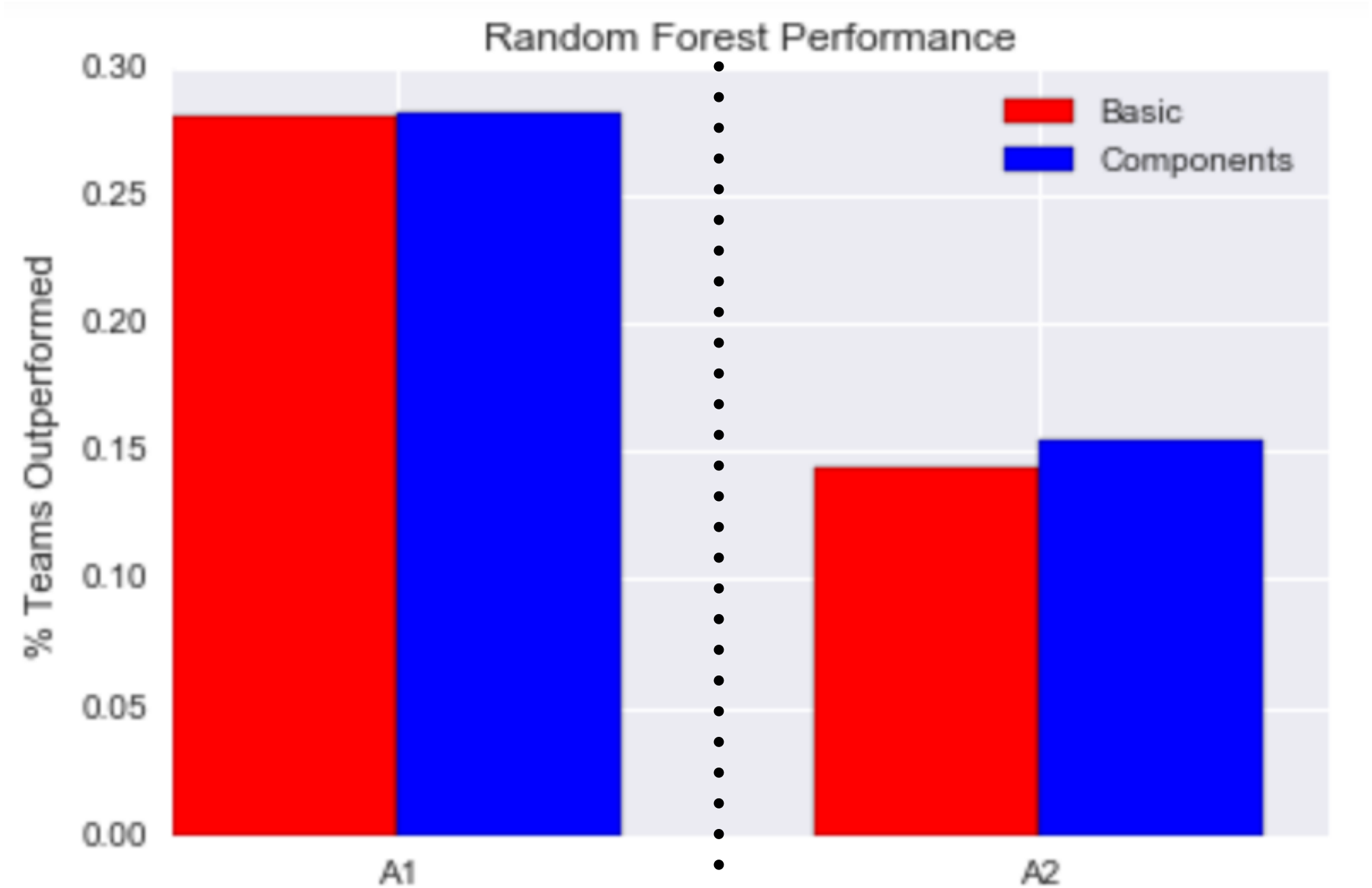
Expectations

1. component > basic
2. GB > RF > DT

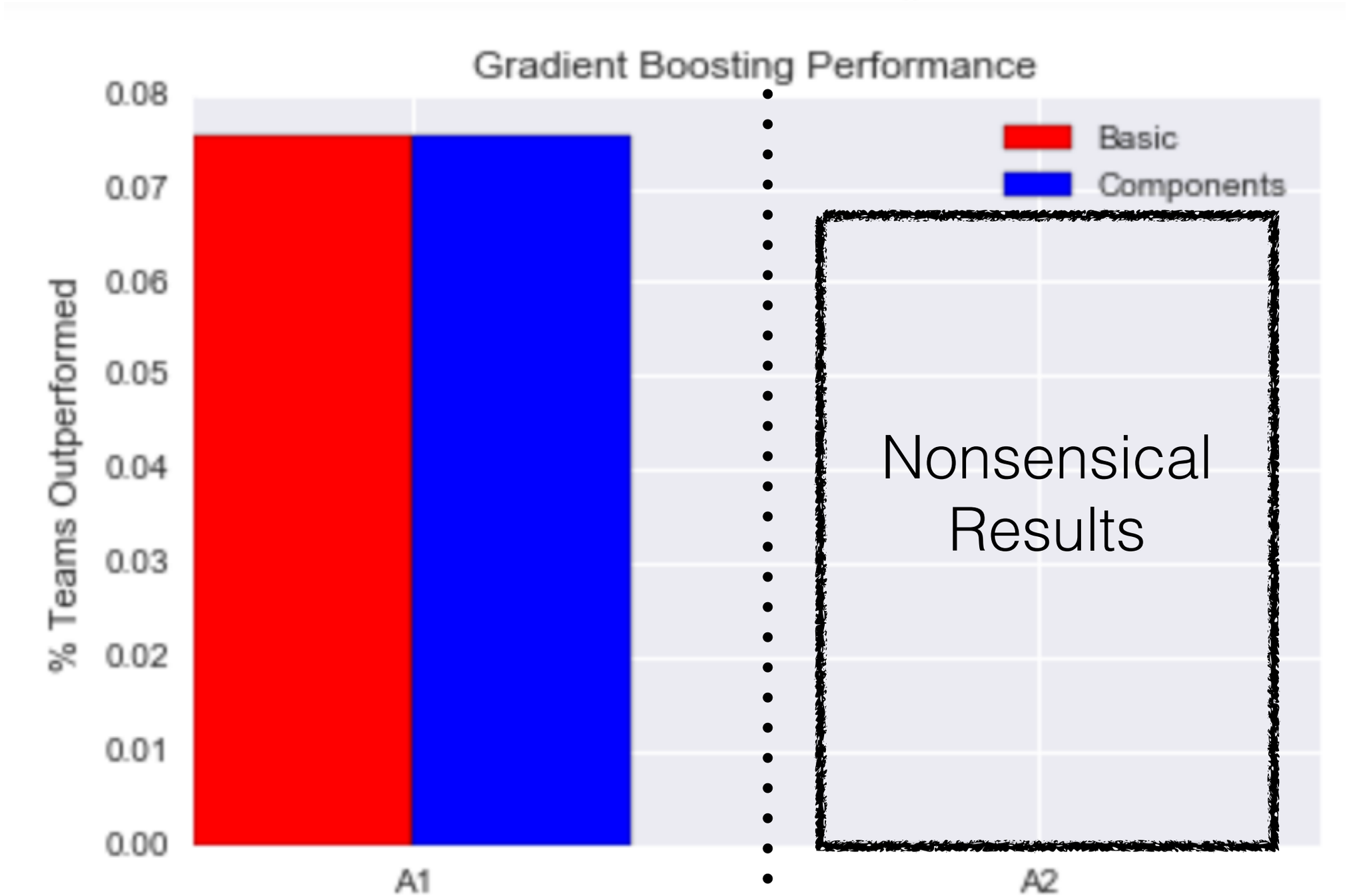
Modeling


Decision Tree Performance
% Teams Outperformed
0.30
0.25
0.20
0.15
0.10
0.05
0.00
A1
A2
A3
A4
Basic
Components

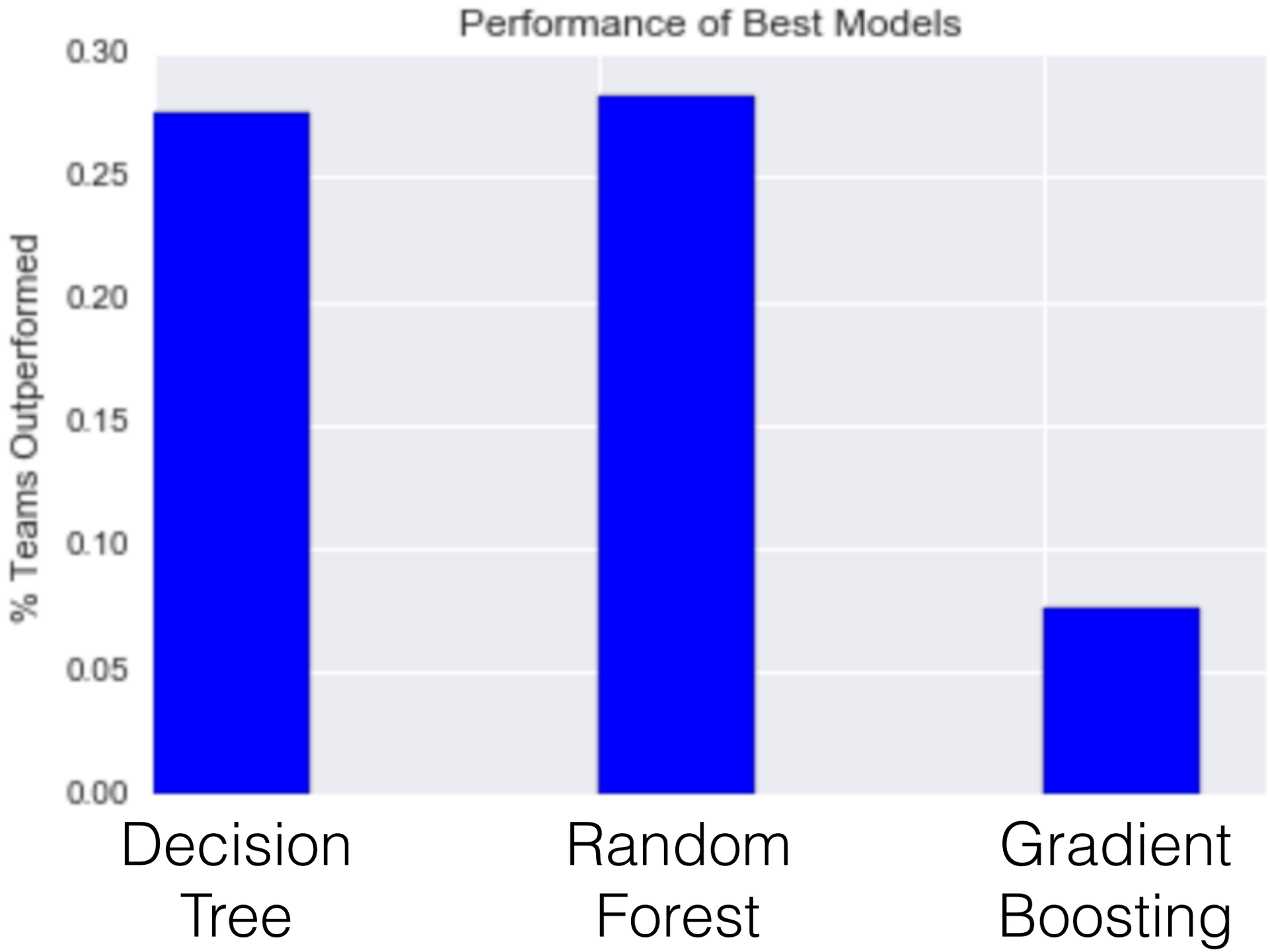
Modeling


Random Forest Performance
0.30
0.25
0.20
0.15
0.10
0.05
0.00
% Teams Outperformed
Basic
Components
A1
A2

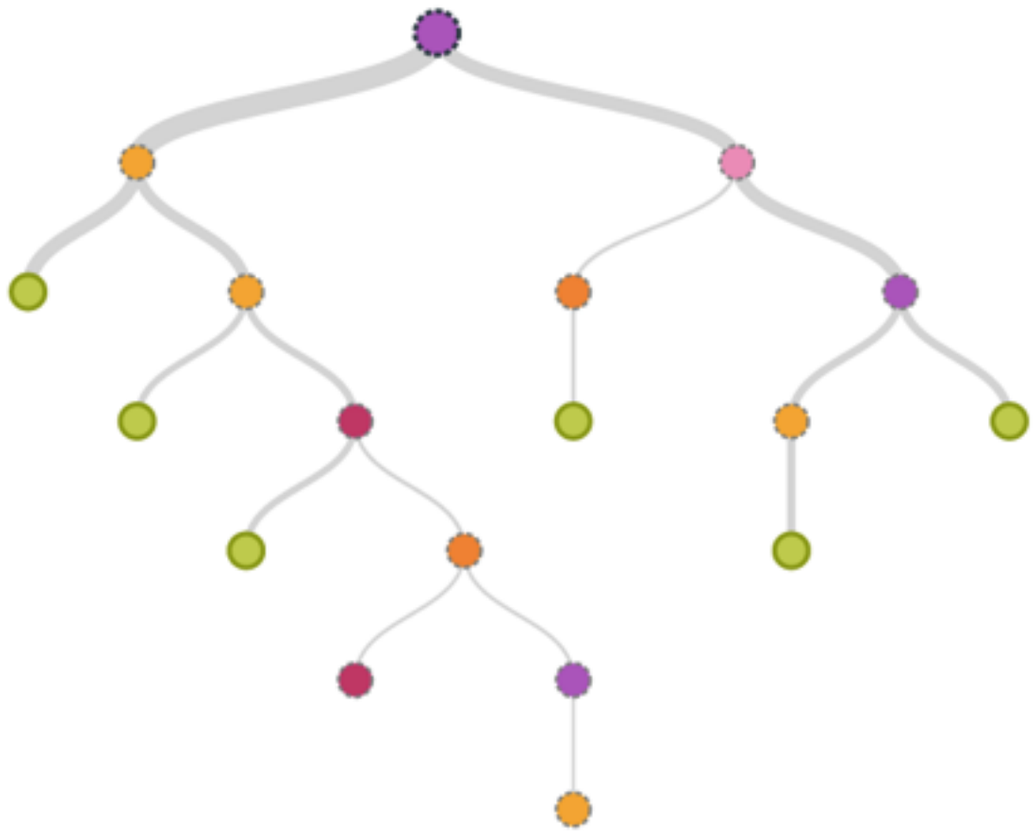
Modeling


Gradient Boosting Performance
Basic
Components
% Teams Outperformed
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
A1
A2
Nonsensical Results

Modeling


Performance of Best Models
% Teams Outperformed
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Decision Tree
Random Forest
Gradient Boosting

Modeling



Expectations

1. component > basic
2. GB > RF > DT

Lessons

Lessons

- more data didn’t clearly improve performance

Lessons

- more data didn’t clearly improve performance
- RF, GB didn’t clearly outperform DT

Lessons

- more data didn’t clearly improve performance
- RF, GB didn’t clearly outperform DT
- tuning parameter issues:

Lessons

- more data didn’t clearly improve performance
- RF, GB didn’t clearly outperform DT
- tuning parameter issues:
  - which parameters to tune?

Lessons

- more data didn’t clearly improve performance
- RF, GB didn’t clearly outperform DT
- tuning parameter issues:
  - which parameters to tune?
  - consequences of suboptimal parameter values

Lessons

- more data didn’t clearly improve performance
- RF, GB didn’t clearly outperform DT
- tuning parameter issues:
    - which parameters to tune?
    - consequences of suboptimal parameter values
- importance of logging tuning runs and model fits

Improvements

Improvements

- more detailed and diligent logging

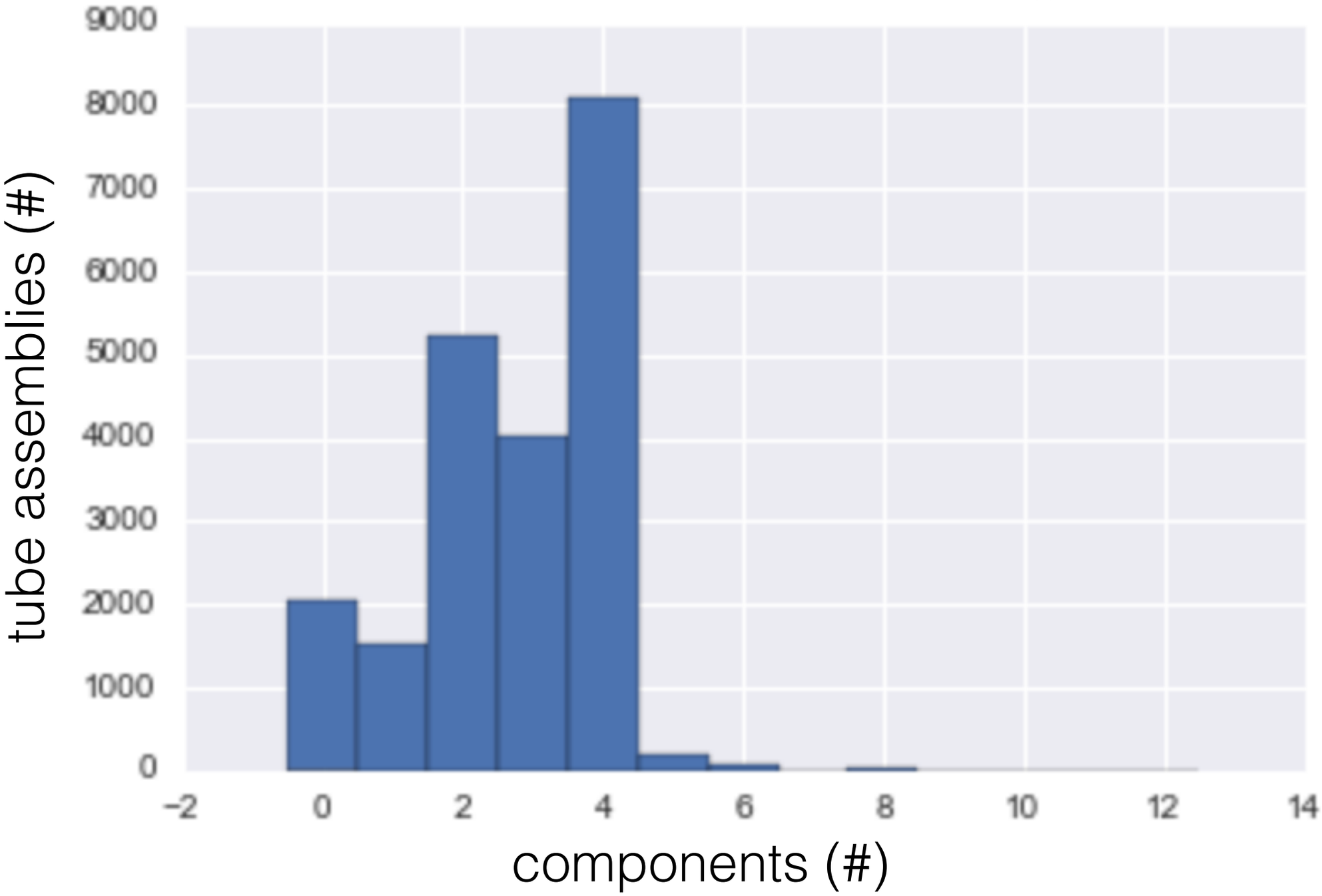
Improvements

- more detailed and diligent logging
- tree splits and parameter selection based on Kaggle metric

Improvements

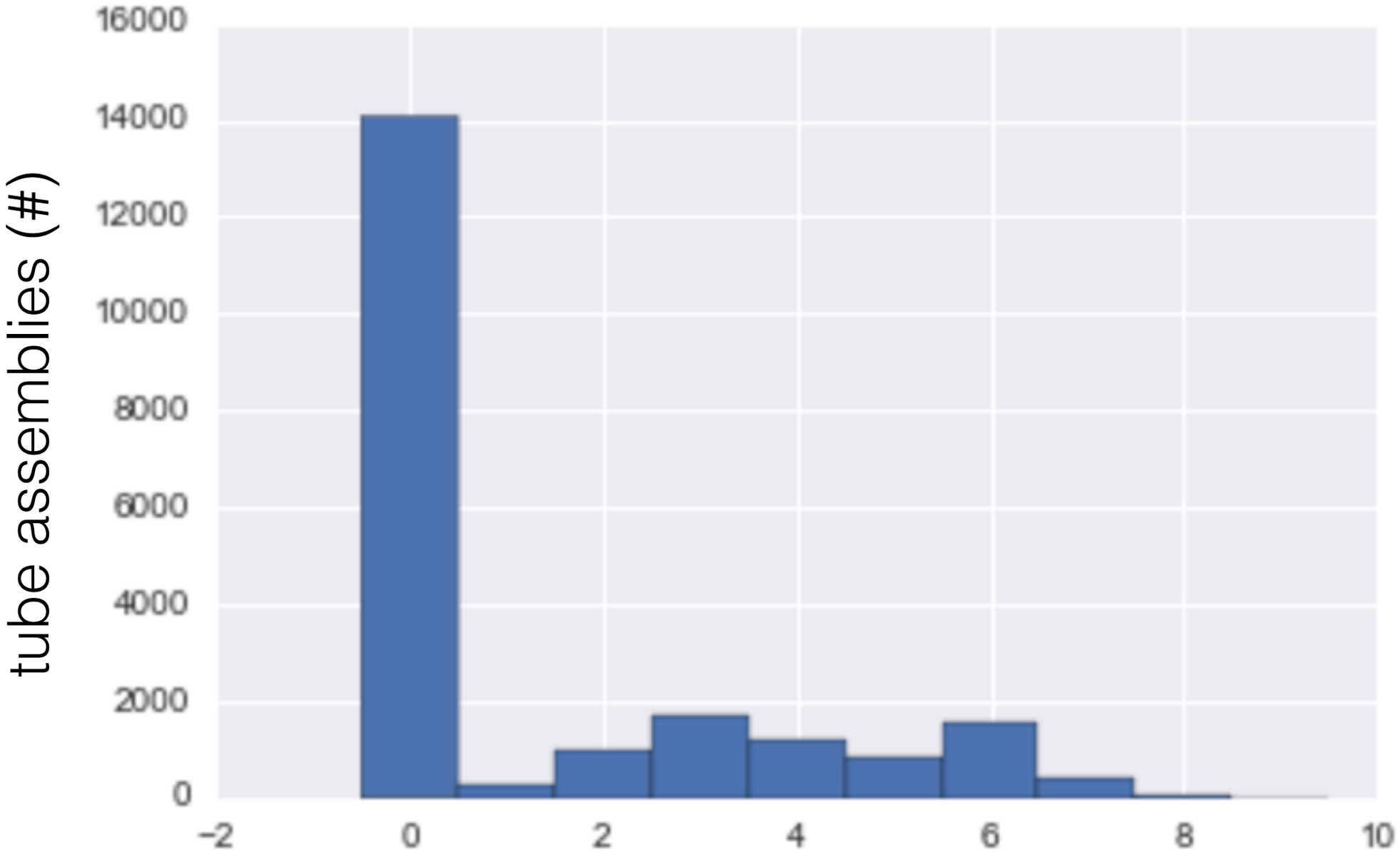
- more detailed and diligent logging
- tree splits and parameter selection based on Kaggle metric
- Spark

Data

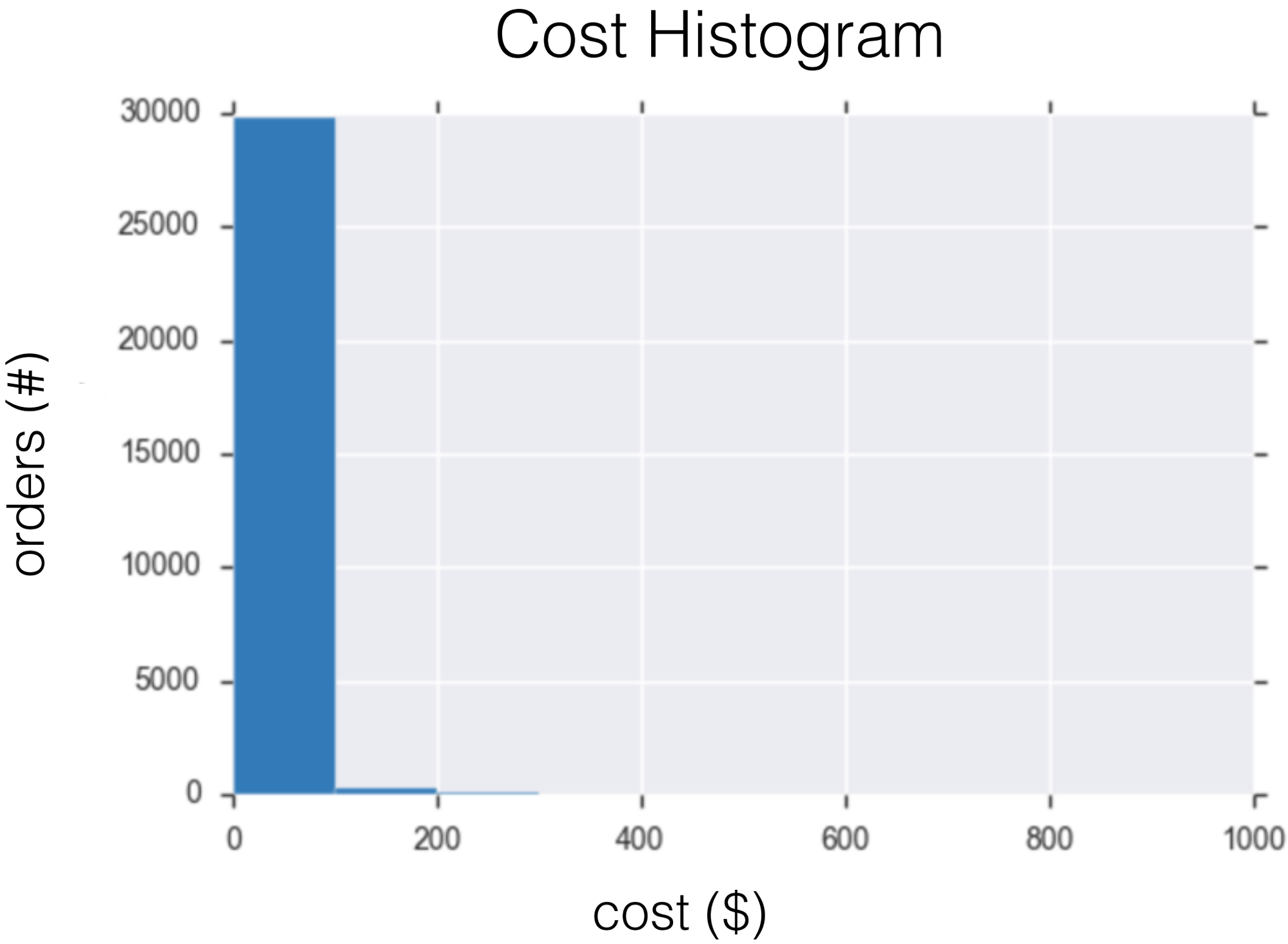

Component Count Histogram
tube assemblies (#)
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
−2
0
2
4
6
8
10
12
14
components (#)

Data

Specifications Count Histogram



Data


Cost Histogram
30000
25000
20000
15000
10000
5000
0
orders (#)
0
200
400
600
800
1000
cost ($)

Data

	tube_assembly_id	supplier	quote_date	annual_usage	min_order_quantity	bracket_pricing	quantity	cost
0	TA-00002	S-0066	2013-07-07	0	0	Yes	1	21.905933
1	TA-00002	S-0066	2013-07-07	0	0	Yes	2	12.341214
2	TA-00002	S-0066	2013-07-07	0	0	Yes	5	6.601826
3	TA-00002	S-0066	2013-07-07	0	0	Yes	10	4.687770
4	TA-00002	S-0066	2013-07-07	0	0	Yes	25	3.541561
5	TA-00002	S-0066	2013-07-07	0	0	Yes	50	3.224406
6	TA-00002	S-0066	2013-07-07	0	0	Yes	100	3.082521
7	TA-00002	S-0066	2013-07-07	0	0	Yes	250	2.999060
8	TA-00004	S-0066	2013-07-07	0	0	Yes	1	21.972702
9	TA-00004	S-0066	2013-07-07	0	0	Yes	2	12.407983
10	TA-00004	S-0066	2013-07-07	0	0	Yes	5	6.668596
11	TA-00004	S-0066	2013-07-07	0	0	Yes	10	4.754539
12	TA-00004	S-0066	2013-07-07	0	0	Yes	25	3.608331
13	TA-00004	S-0066	2013-07-07	0	0	Yes	50	3.291176

Decision Tree Tuning Basic

Attempt	min_samples_leaf	min_samples_split	test_size	max_leaf_nodes
A1	2	2	0.20	-
A2	2	23.3	0.01	-
A3	3	7.3	0	-
A4	-	-	0	280

Decision Tree Tuning Components

Attempt	min_samples_leaf	min_samples_split	test_size	max_leaf_nodes
A1	4	2	0.20	-
A2	4	23.3	0.01	-
A3	2	18	0	-
A4	-	-	0	9

Random Forest Tuning Basic

Attempt	min_samples_leaf	min_samples_split	test_size	n_estimators	max_leaf_nodes
A1	1	2	0	100	-
A2	-	-	0	1000	700

Random Forest Tuning Components

Attempt	min_samples_leaf	min_samples_split	test_size	n_estimators	max_leaf_nodes
A1	1	2	0	100	-
A2	-	-	0	1000	650

Gradient Boosting Tuning Basic

Attempt	learning_rate	min_samples_split	test_size	n_estimators	max_leaf_nodes
A1	0.01	-	0	100	2

Gradient Boosting Tuning Components

Attempt	learning_rate	min_samples_split	test_size	n_estimators	max_leaf_nodes
A1	0.01	-	0	100	2