Predicting Retention

WENQI CHENG
MARIO DIAZ DE LA ROSA
CLARK LI

Cleaning up the Data - Target

Retained User = Active 30 days prior to data pull.

Data pulled on July 1st, 2014.

Consider user retained whenever

last_trip_date >= '2014-06-01'

This defines our classification labels → find factors that best predict value of **1**

All features are arguably, intuitively relevant, so we decided to evaluate them all.

Cleaning up the Data - Numbers

Additional engineering:

- Dummified phone (Android = 1, iPhone = 0)
- Dummified city (King’s Landing and Winterfell)
- Created an Account_Age metric by calculating number of days from signup_date to ***data pull***
    - ***NOT*** to last_trip_date to prevent leakage

Cleaning up the Data - Ratings

The Ratings Question:

- *what do the NaN represent?*
- *are we losing a valuable signal by removing NaN data?*
- *can the way customers use ratings inform our decision?*

Trade-off between potential information loss and additional data.

Cleaning up the Data – Douchebag Matrix

Uber's 'douchebag rating' is still a source of shame for riders

2.4k
SHARES

Share on Facebook | Share on Twitter

WHAT'S THIS?

IMAGE: MASHABLE/VICKY LETA

BY SETH FIEGERMAN
OCT 02, 2015

Daniela Cadena @DanielaCadena — Follow

My Uber rating is 4.7 and I'm kind of bummed I'm not a 5. Where did I go wrong?

5:16 PM - 21 Sep 2015

♥ 2

casse-toi @camiejuan — Follow

My Uber rating is 4.9. I wonder what I did to lose the 5 rating. Probably that time I ate my baon in the car and it stank. But I apologized!

9:38 PM - 22 Sep 2015

♥ 29

Emmy Jo Favilla @em_dash3 — Follow

My uber rating is a 4.4 and I'm feeling pretty salty about that 0.6

5:07 PM - 21 Sep 2015

♥ 1

Alex Miller @AlexTheMiller — Follow

Grad School: Hey what's your GPA
Me: Well my Uber rating is a 5 and I think that says a lot

8:08 AM - 29 Sep 2015

13 ♥ 42

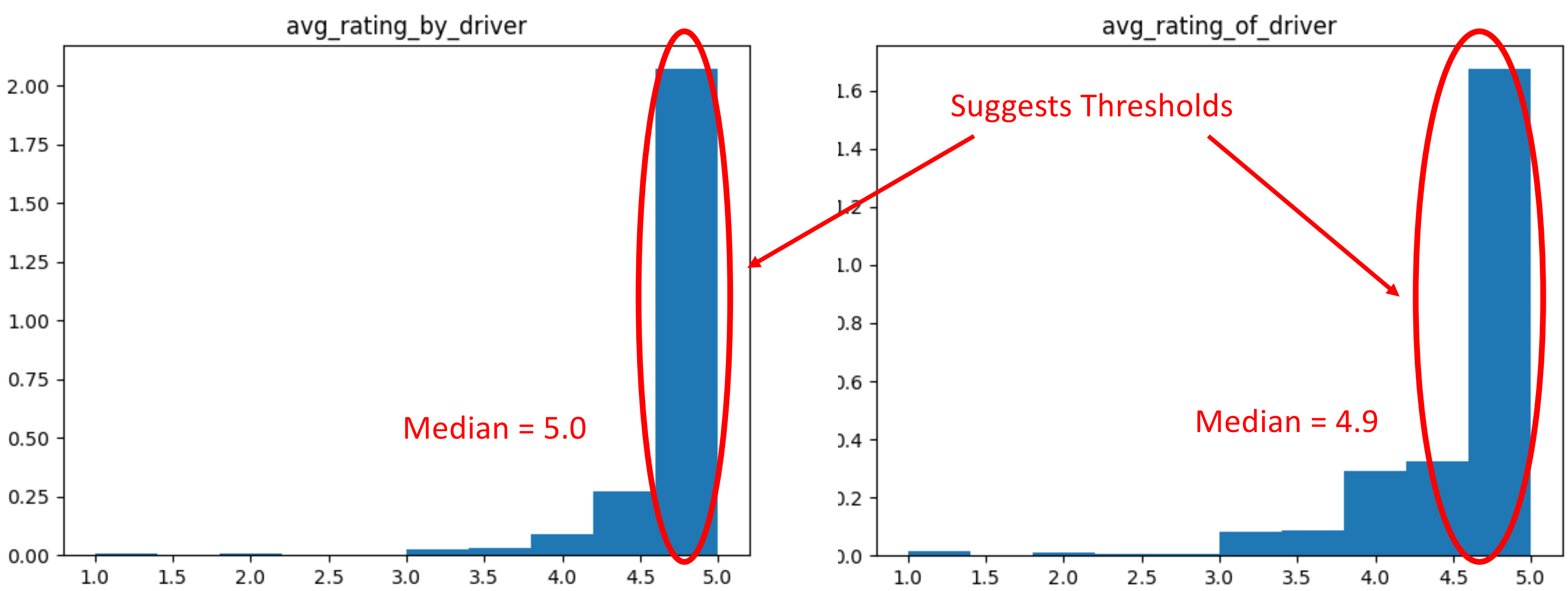
jon bassett @jon_bassett — Follow

@Uber drivers taking their ratings seriously

9:05 AM - 9 Mar 2015

2 ♥ 5

Cleaning up the Data - Ratings



Naïve Bayes

Classifier applies Bayes Theorem assuming independence between features.

For the validation data set:

Multinomial NB accuracy = 68.9%

Gaussian NB accuracy = 73.8%

Not the best model to use due to low accuracy of MNB and difficult to interpret of justify GNB

Logistic Regression vs. Random Forest vs. Gradient Boosting

Logistic Regression		Predicted	
		Not Churn	Churn
Actual	Not churn	2217	1555
	Churn	1030	5198

Accuracy

Train	0.75
Test	0.74

Precision: 0.68

Recall: 0.59

Random Forest		Predicted	
		Not Churn	Churn
Actual	Not churn	2370	1402
	Churn	867	5361

Train	0.81
Test	0.77

Precision: 0.73

Recall: 0.63

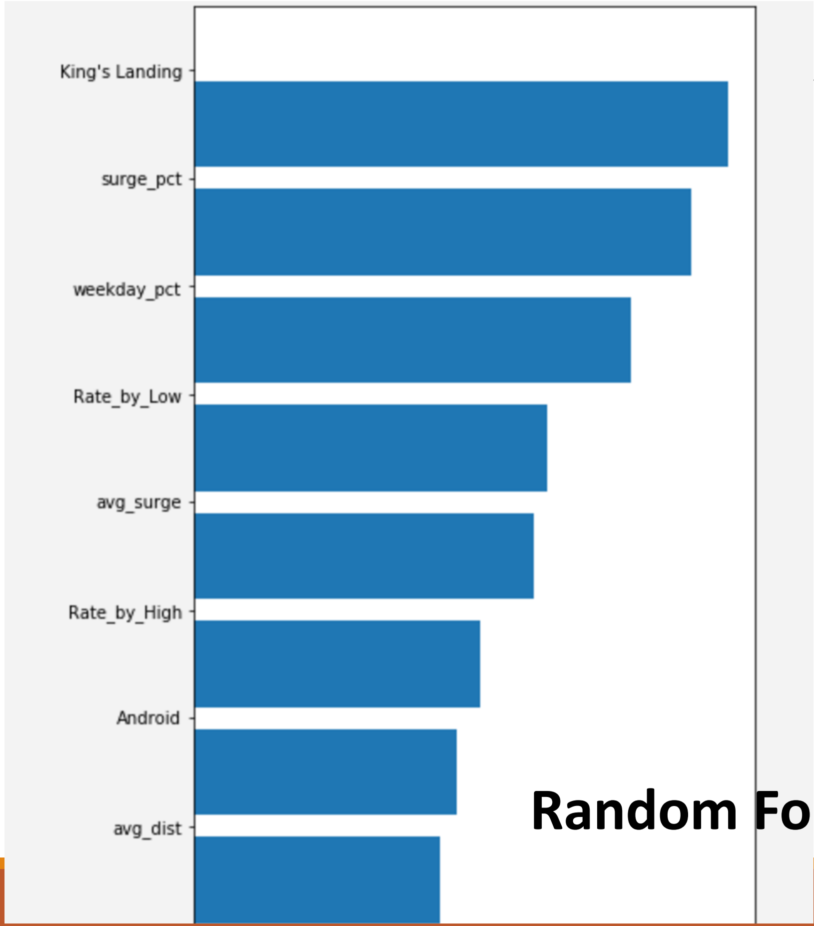
Gradient Boosting		Predicted	
		Not Churn	Churn
Actual	Not churn	2408	1364
	Churn	883	5345

Train	0.79
Test	0.77

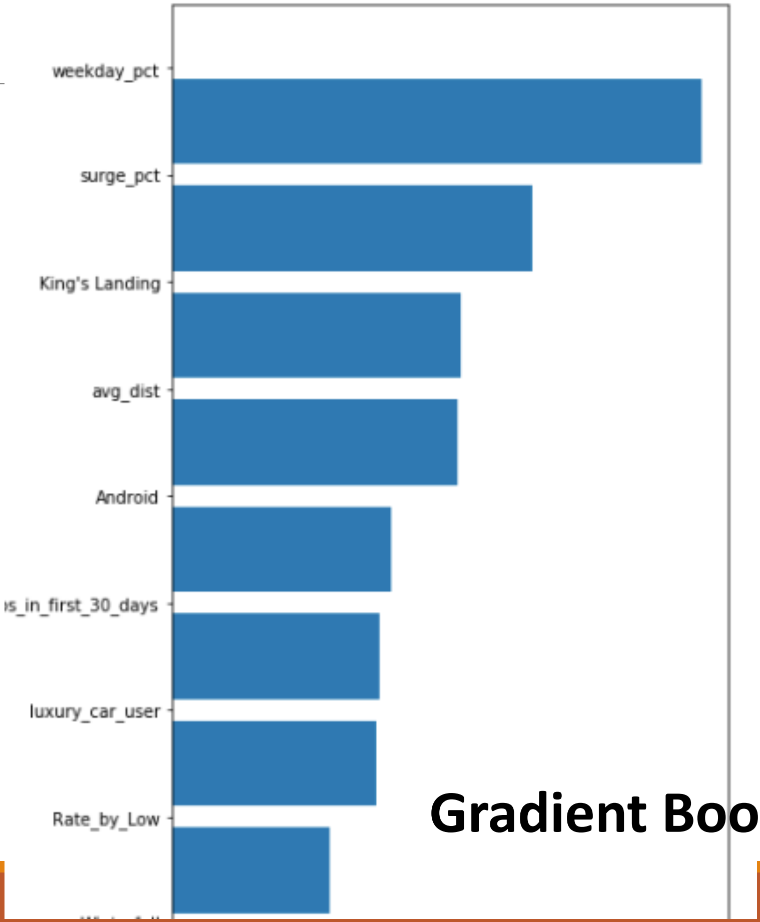
Precision: 0.73

Recall: 0.64

Important Features



Random Forest



Gradient Boosting

Conclusion & Future Work

- Random Forest and Gradient Boosting gave better predictions

- Focus on:
  - Weekday Trip Percentage
  - Percentage of Trips with Surge Multiplier > 1

- Further validation of our models is necessary
- Need more data that reflects broader distribution of users

Team Kuma

Questions?