Google

official merchandise store

Customer Revenue Prediction

UC Berkeley
Graduate Data Science Organization

Data Science Workshop 2019

Mentor
Andy Vargas

Team
Yuem Park
Marvin Pohl
Michael Yeh

Why do we care?

"

The 80/20 rule has proven true for many businesses - **only a small percentage of customers produce most of the revenue**. As such, marketing teams are challenged to make appropriate investments in promotional strategies.

...

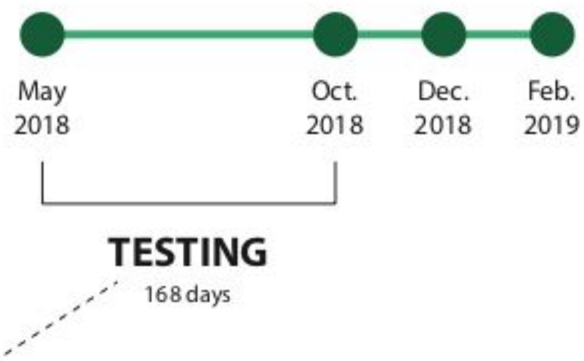
Hopefully, the outcome will be more **actionable operational changes and a better use of marketing budgets** for those companies who choose to use data analysis on top of Google Analytics data.

"

- Kaggle competition website

Project Overview

Project Overview

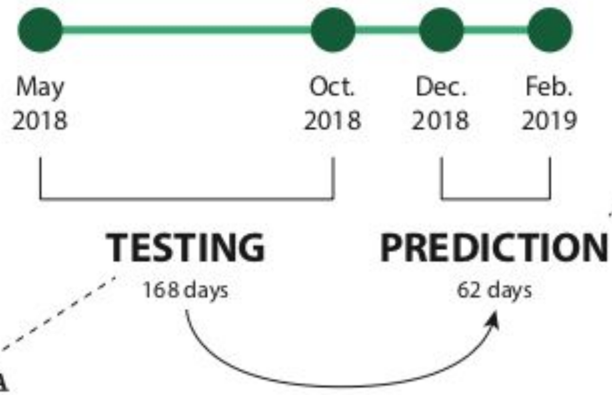


DATA

Each row: one visit to the store

Each column: data about that visit:
- visitor ID
- date
- device
- location
- time spent on website
- revenue generated
- many, many more...

Project Overview



PREDICTION

Asked to predict how much revenue customers who visited the store between May and October 2018 are GOING to generate between December 2018 and February 2019.

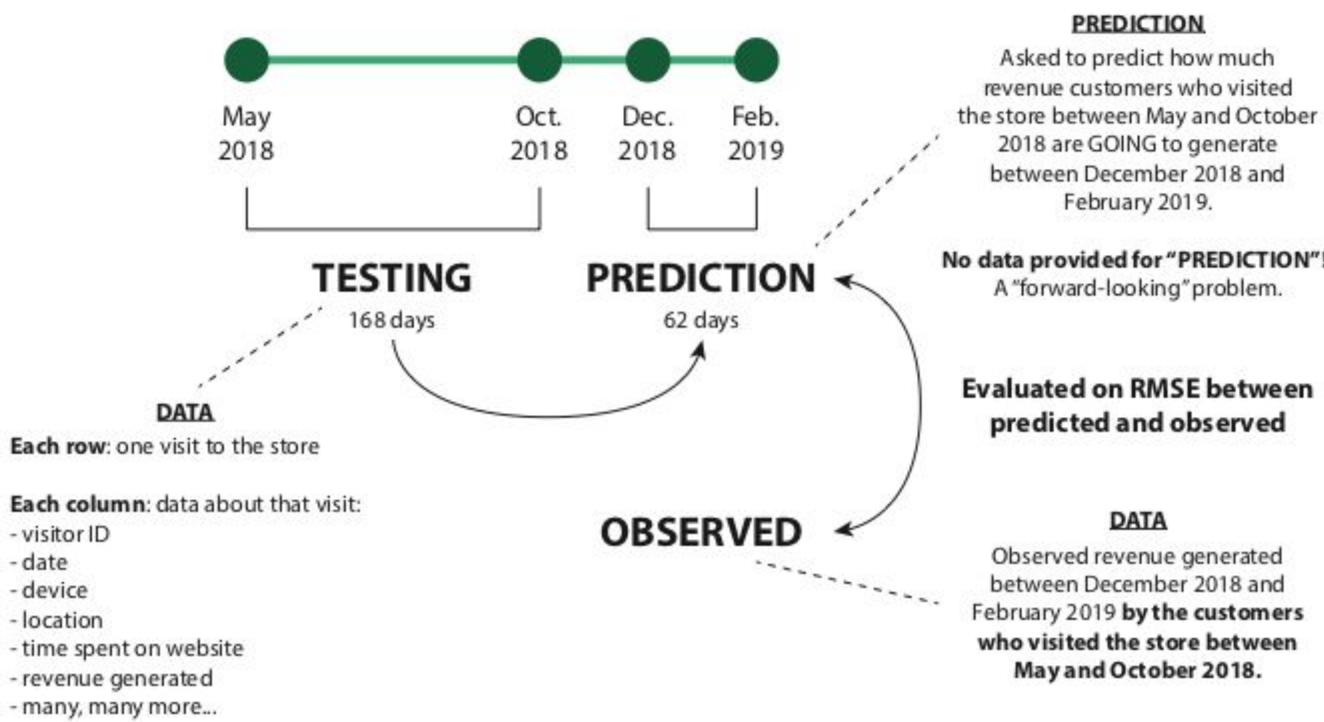
No data provided for "PREDICTION"!!
A "forward-looking" problem.

DATA

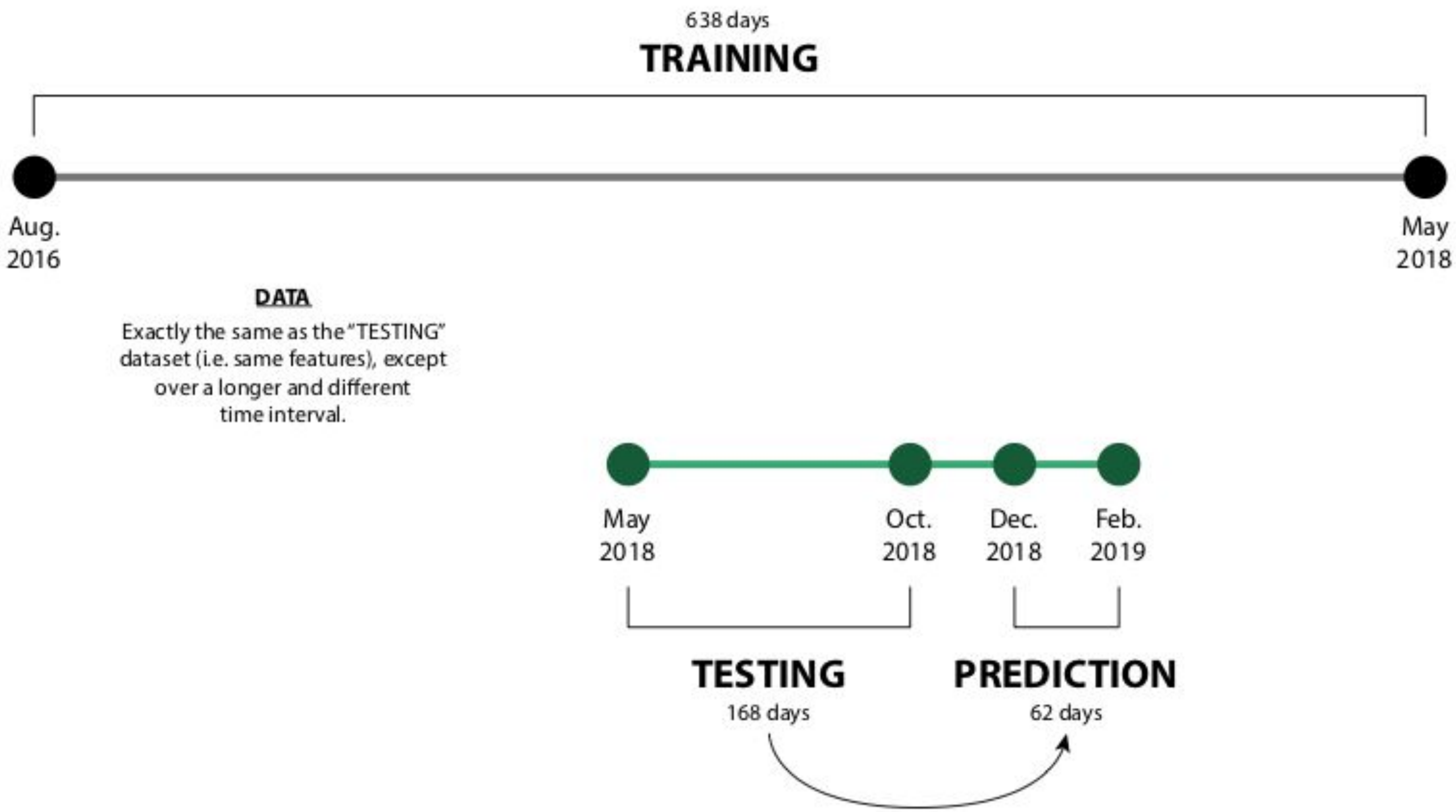
Each row: one visit to the store

Each column: data about that visit:
- visitor ID
- date
- device
- location
- time spent on website
- revenue generated
- many, many more...

Project Overview

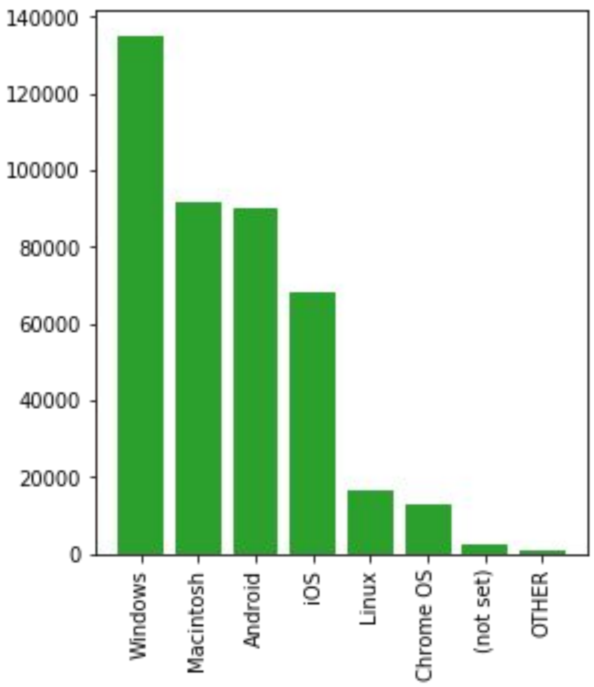


Project Overview

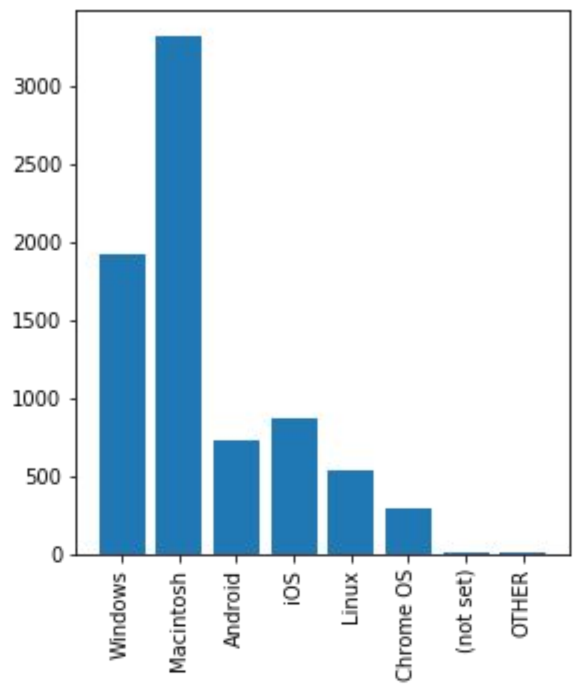


Exploratory data analysis

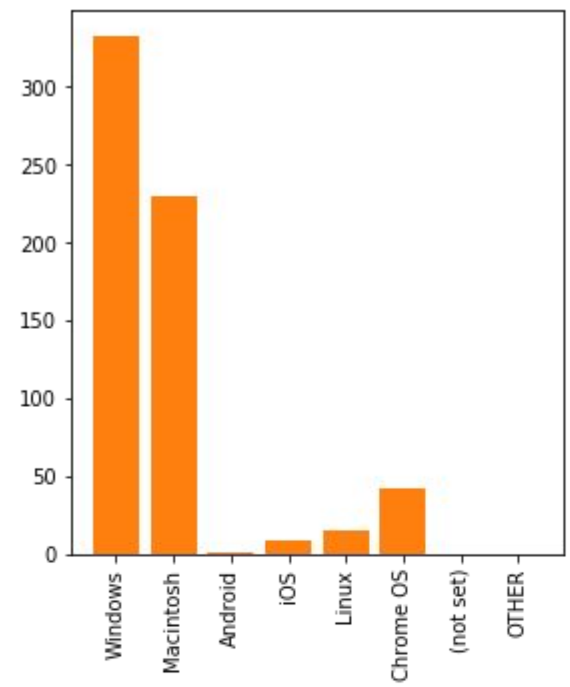
Did not return



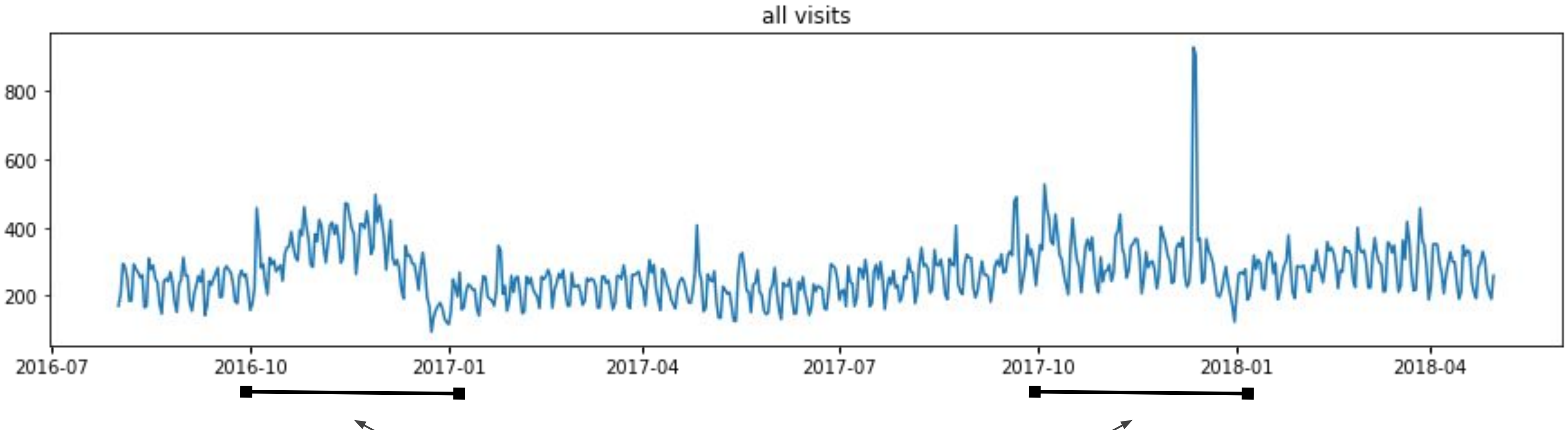
Returned, didn’t spend



Returned and spent

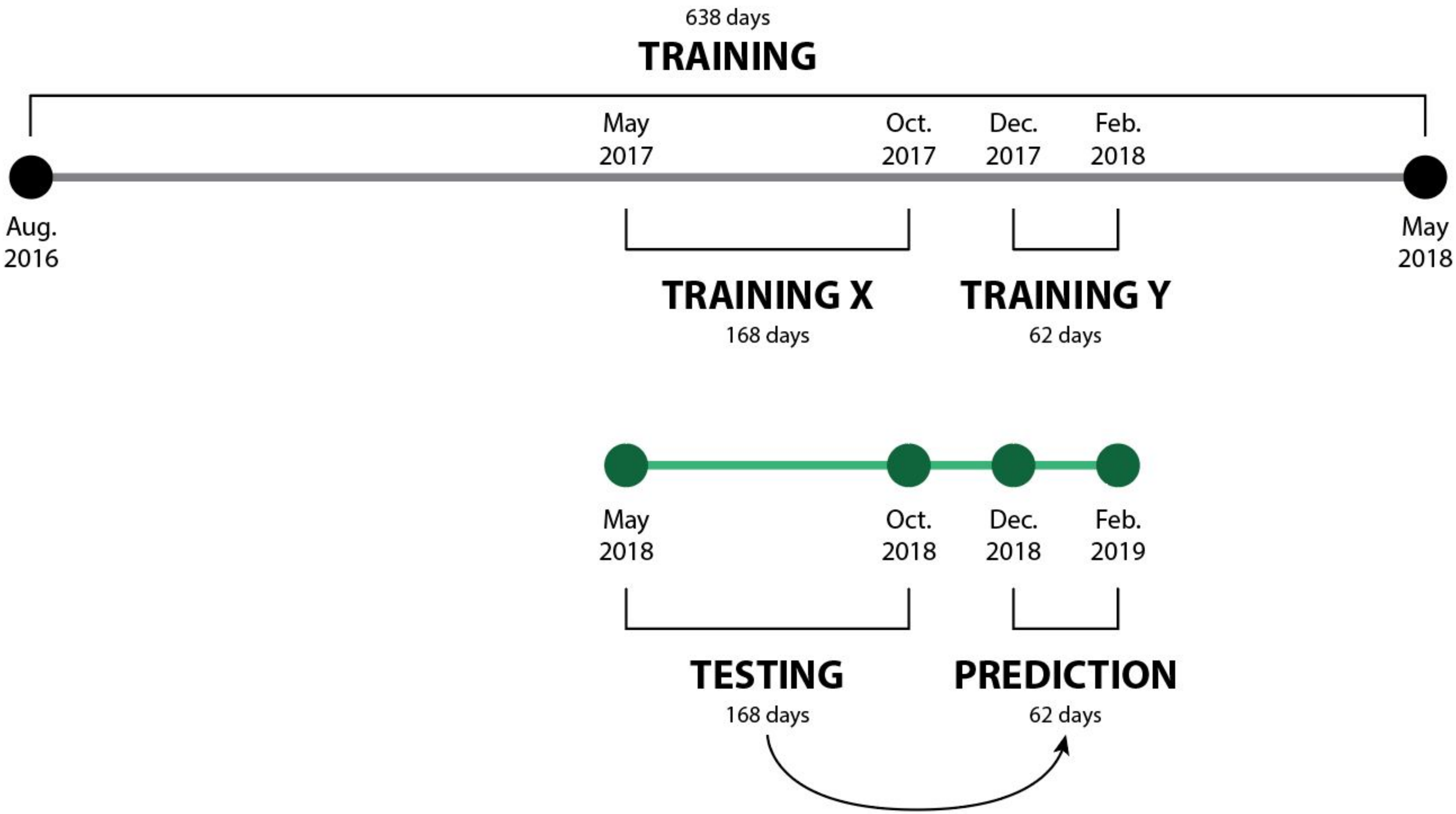


Seasonality

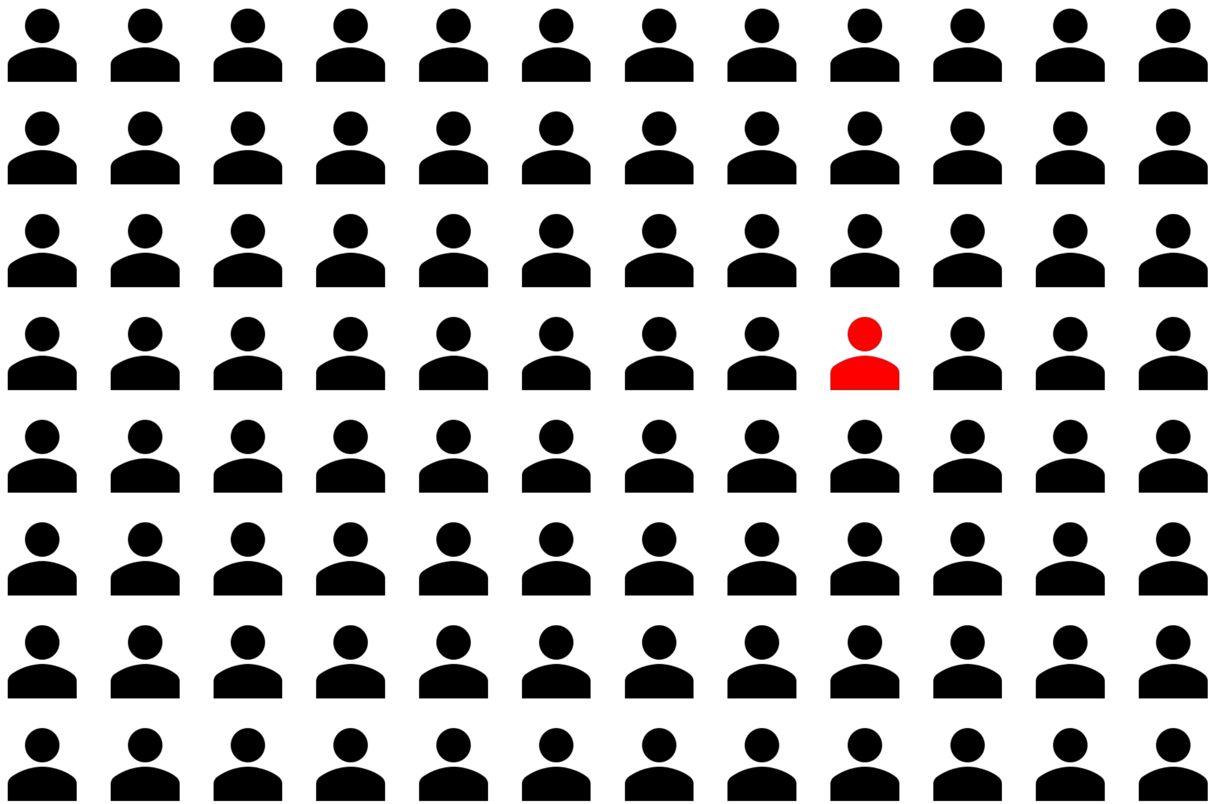


More visits during holiday season

Seasonality

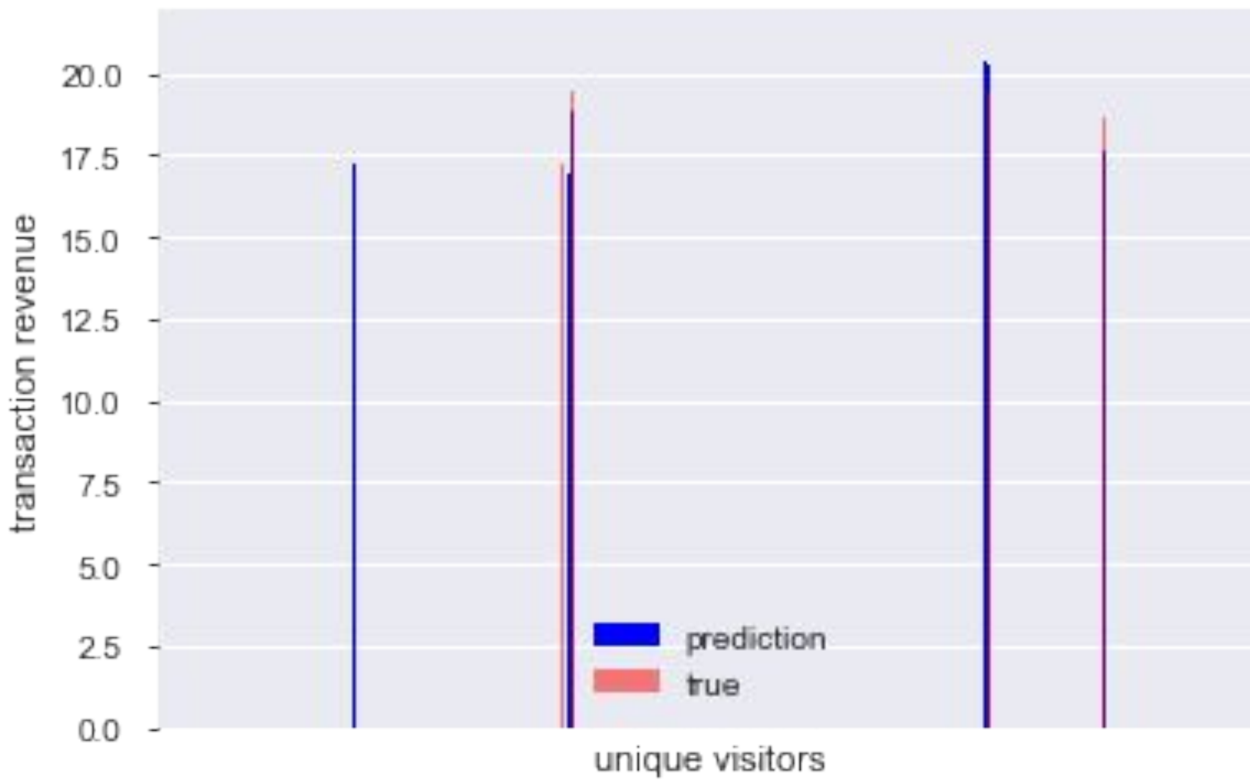


Imbalanced data: < 0.04% returned and spent!



Imbalanced data: < 0.04% returned and spent!

RFR on all visitors that appear in both TRAINING X and TRAINING Y



Choosing the right approach

Two-step approach:

1

Using a **logistic regression** to calculate the **probability that a visitor returns & spends**

P

x

2

Using a **random forest regression** to calculate the **total revenue of each visitor**

$

Results

Evaluation criteria:

Root mean square error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

#		Team Name	Team Members	Score	Entries	Last
1	▲ 68	ML Keksika		0.88140	5	9mo
2	▲ 31	pika pika pikachu		0.88202	8	9mo
3	▲ 905	zxasd131		0.88273	2	9mo
...						
147		**BASELINE = [0,0,0,...,0]**		**0.88843**		

(out of 1089 teams)

Results

<u>Evaluation criteria:</u>

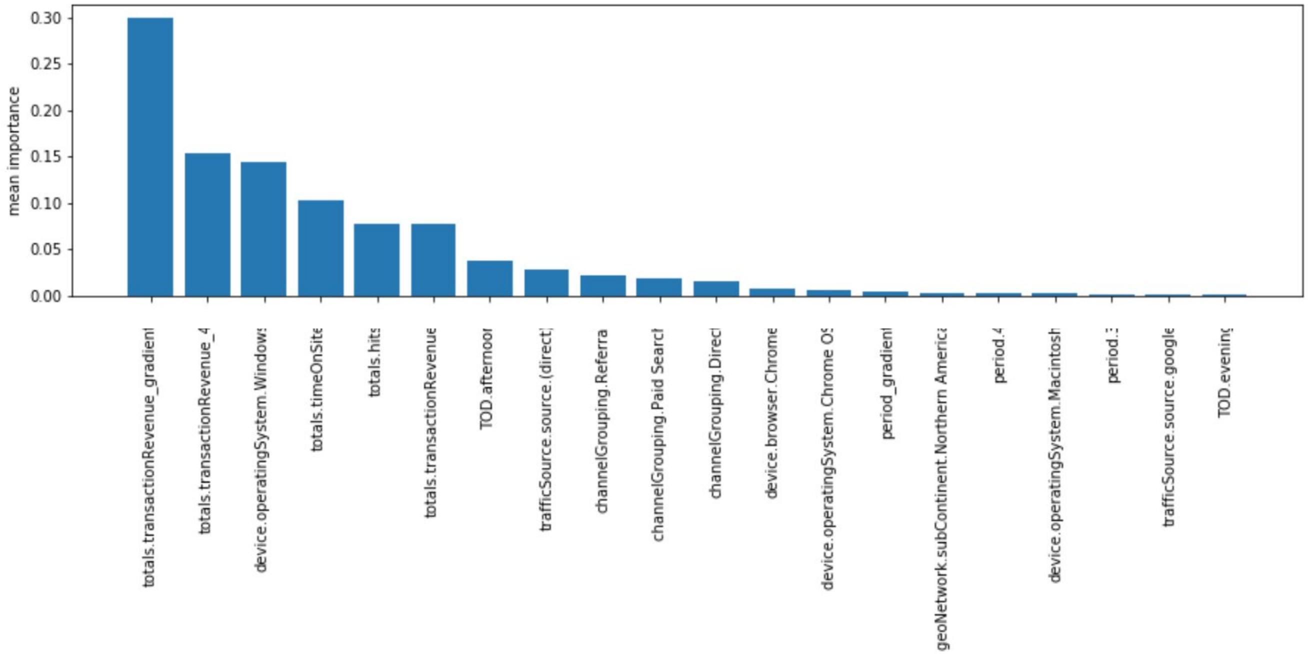
Root mean square error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2},$$

#		Team Name	Team Members	Score	Entries	Last
1	▲68	**ML Keksika**		0.88140	5	9mo
2	▲31	**pika pika pikachu**		0.88202	8	9mo
3	▲905	**zxasd131**		0.88273	2	9mo
...						
34		**DSW**		**0.88516**		
...						
147		**BASELINE = [0,0,0,...,0]**		**0.88843**		

(out of 1089 teams)

Feature importance



<u>To answer the entry question:</u>

Our recommendation is to focus marketing budgets into those visitors that:

- **Spent in the past**
  (preferentially with positive gradient)
- **Use Windows OS**
- **Are recurring visitors**
- **Visit in the afternoon**

Final thoughts and takeaways

- Carefully selected features and creating new features was key
- More actual spenders would have enabled us to use more accurate but data-heavy models (e.g. NN)

- If data is heavily unbalanced it is hard to predict much better than a zero baseline
- The two-model approach predicted a non-zero spending for each user

- Data preprocessing takes about 90% of the time (at least for us)

Thanks, especially Andy and the GDSO organizers