AGENDA

I. QUICK INTRO
II. APPROACH
III. CHALLENGES
IV. NEXT STEPS
V. Q&A

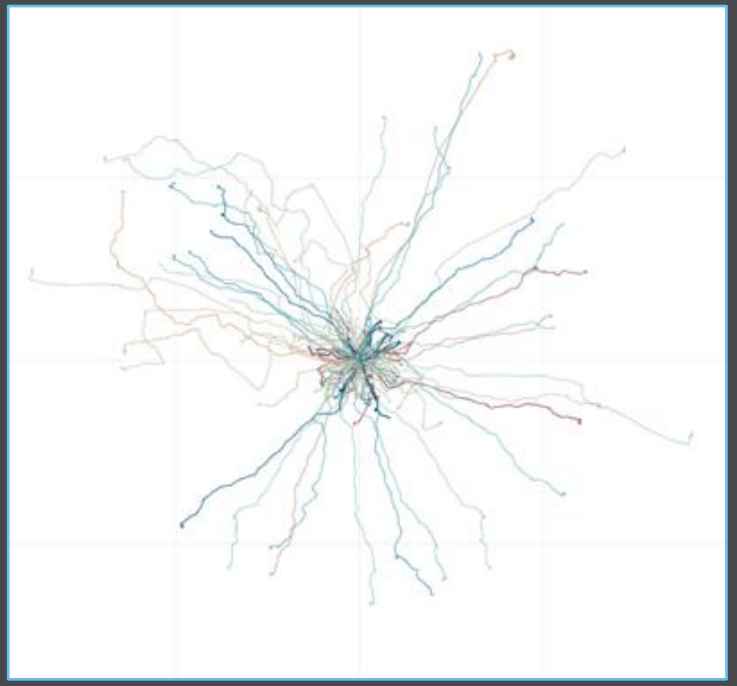
QUICK INTRO

$30,000 *prize*
1,568 *teams*

Driver Telematics Analysis

Sponsor: AXA is a French investment banking firm interested in predicting driving characteristics

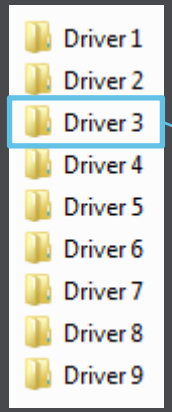
Competition Basics: Use GPS driving data to model a driver signature for 3000+ drivers, then use signatures to identify fake drives inserted for each driver



QUICK INTRO (CONT)

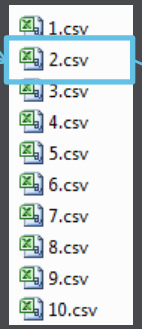
Data

Drivers



3612 total

Drives



200 each

Drive

x	y
0	0
18.6	-11.1
36.1	-21.9
53.7	-32.6
70.1	-42.8
86.5	-52.6
101.7	-62.3
117	-71.6
131.2	-80.4
145.5	-88.7

~1000 each

X and Y coordinates every second

Is this actually Driver 2?

Model

?

Submission

driver_trip	prob
1_1	1
1_2	1
1_3	1
1_4	1
1_5	1
1_6	0
1_7	0
1_8	1
1_9	1
1_10	1
1_11	1
1_12	1
1_13	1

APPROACH

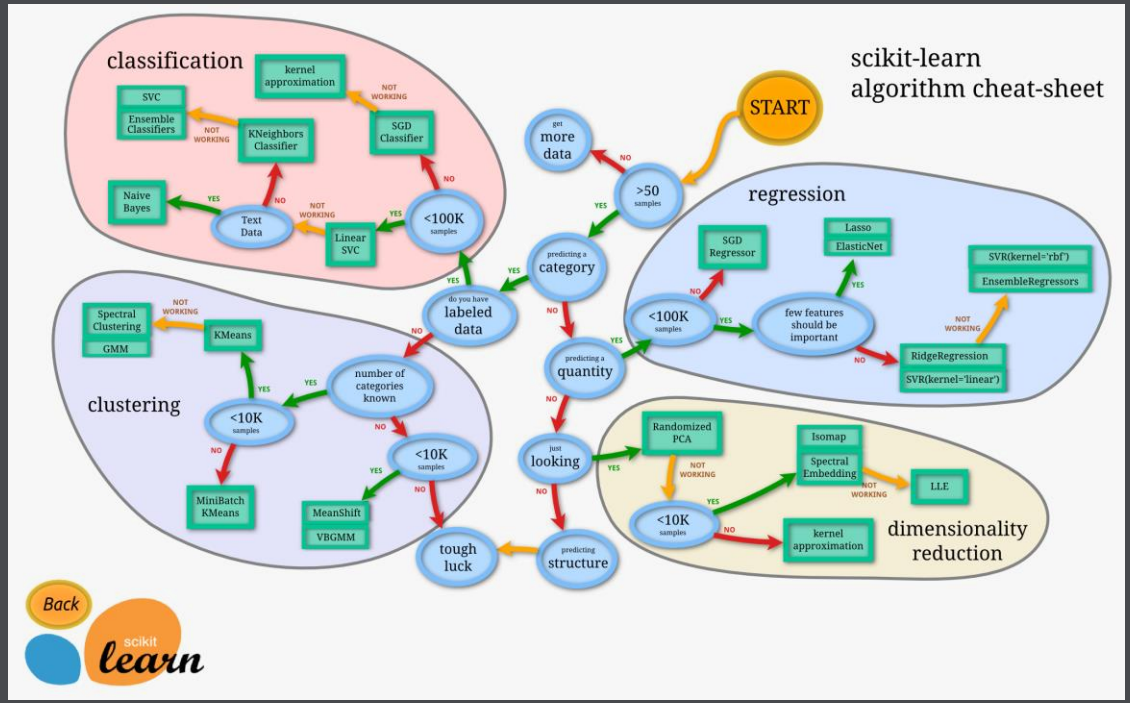
Picking a Model

I have:

- Unlabeled data (K-means?)
- Large-scale, divided data (something simpler?)

I need:

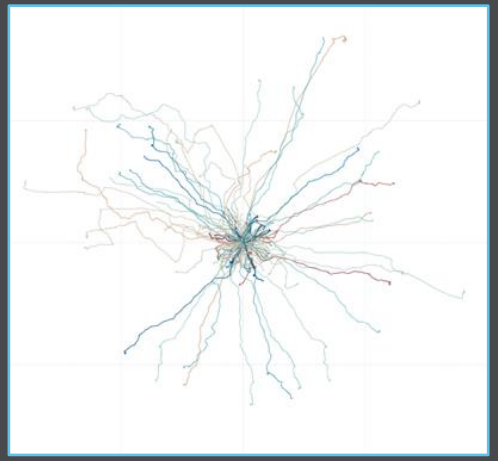
- ‘0’ - ‘1’ predictions (Logistic Reg. or Decision Trees?)
- Can I force labels?



APPROACH (CONT)

Creating Features for the Data

1. Trip Length
2. Average Trip Speed
3. Max Trip Speed
4. Number of turns
5. Degree of turns
6. Acceleration
7. More advanced features…

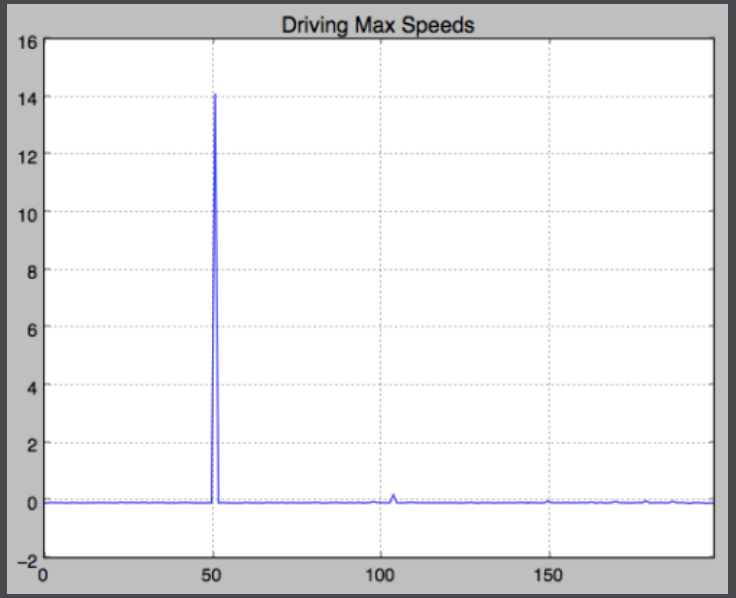


Running the Model

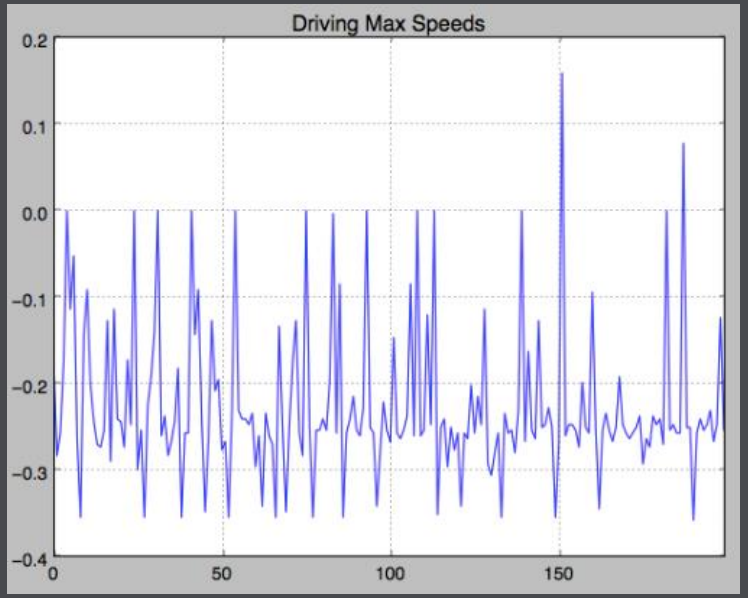
1. Loop through each folder to compile features for each driver
   - Loop through each n+1 driver to grab their features to combine with above driver features
   - Run logistic regression on train-test split data for that driver to get preds
   - Run for all drivers to get preds
2. Combine all driver preds (565k) into one file
3. Submit file to get AUC accuracy ranking
4. Adjust model as necessary, resubmit

CHALLENGES

Do I want to include outliers?

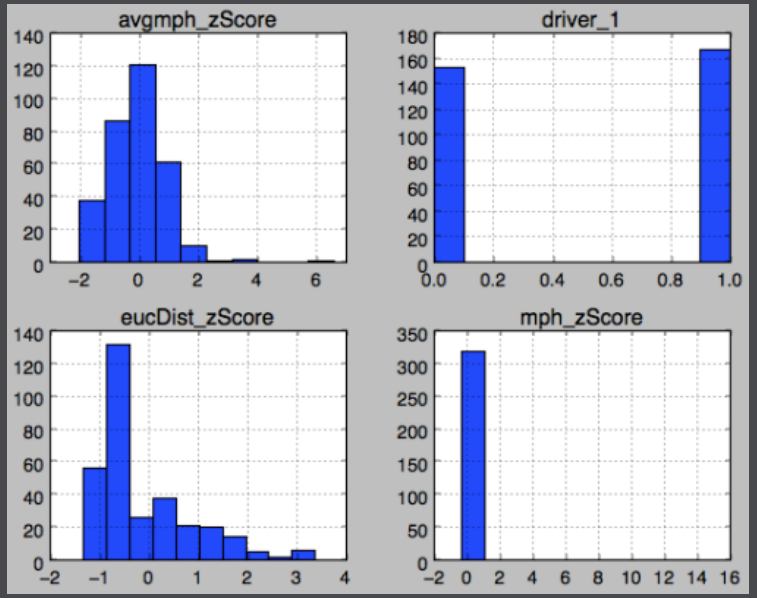


Or Not?



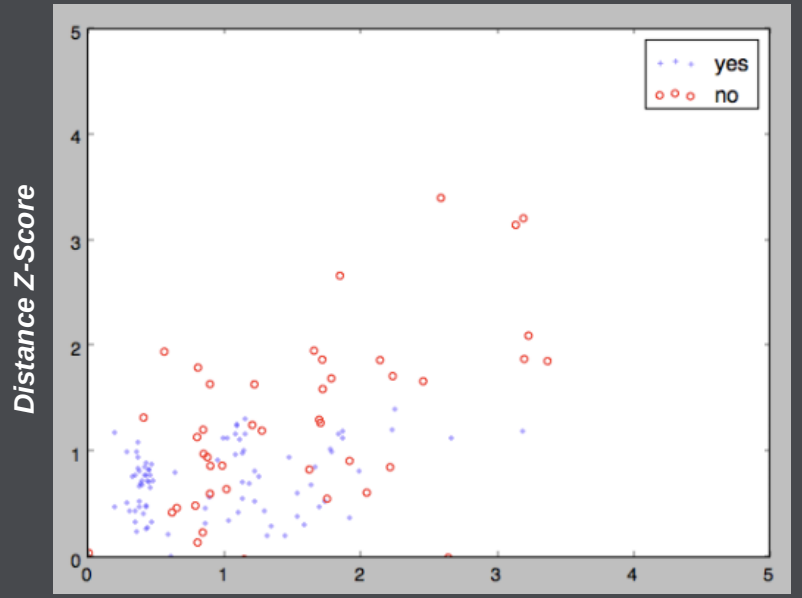
CHALLENGES (CONT)

Features Overview



Overview of Combined Training Set

Is this Driver 1?



CHALLENGES (CONT)

Data Structure and Size

1. **Creating looping structures for folders and files was complicated**
2. **Running my basic model initially took 20 hours**
3. **Introducing more modeling increased this to 30 hours**
4. **Competition submissions required exactly 547,201 rows, mine were always coming up short, not easy to figure out why**

NEXT STEPS

1. Complete initial entry form and submit
2. Review the forums after competition is closed to see winning strategies
3. Keep competing in Kaggle competitions

Q&A