AirBnB Destinations

Tim Rappold
Metis
Project McNulty
Spring 2018

Outline

- ☐ AirBnB Kaggle Challenge Overview
- ☐ Data & Class Imbalance
- ☐ Previous Work & Evaluation Criteria
- ☐ My Results
- ☐ Model selection, how I got my results
- ☐ Conclusions and Next steps

“Sometimes it's the journey that teaches you a lot about your destination.

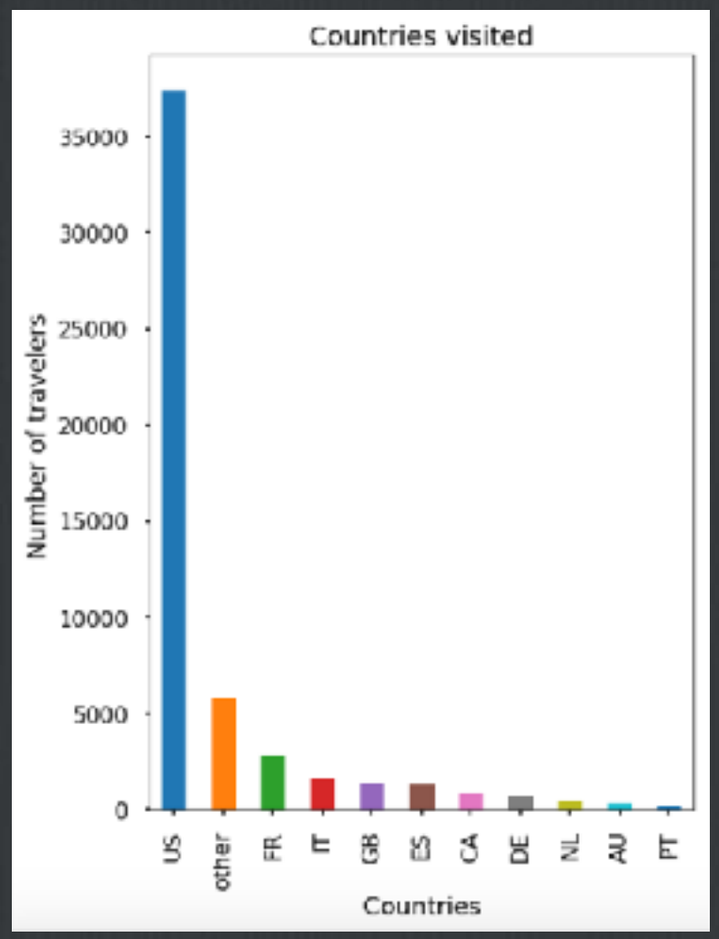
–Drake

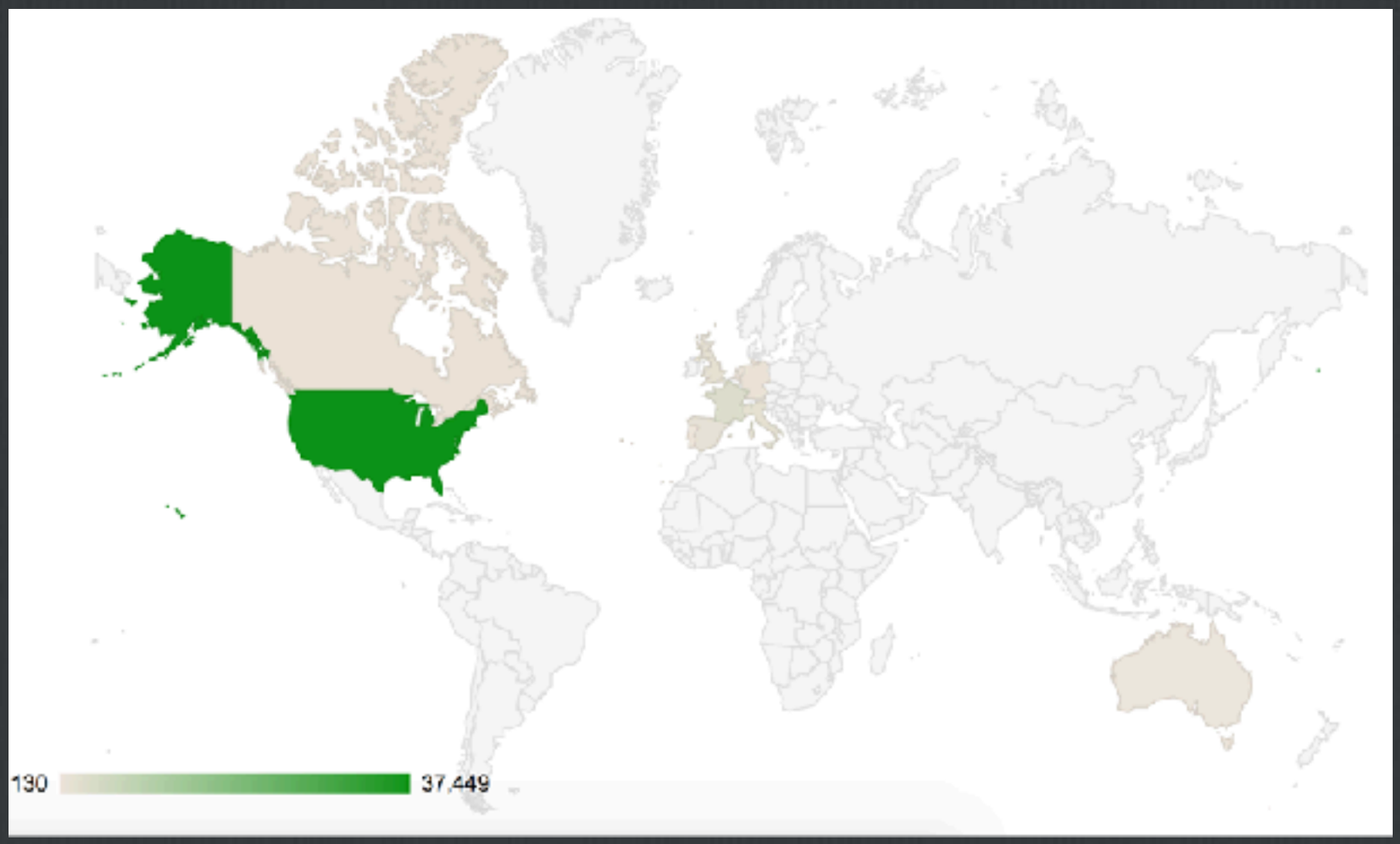
kaggle

AirBnB Destinations

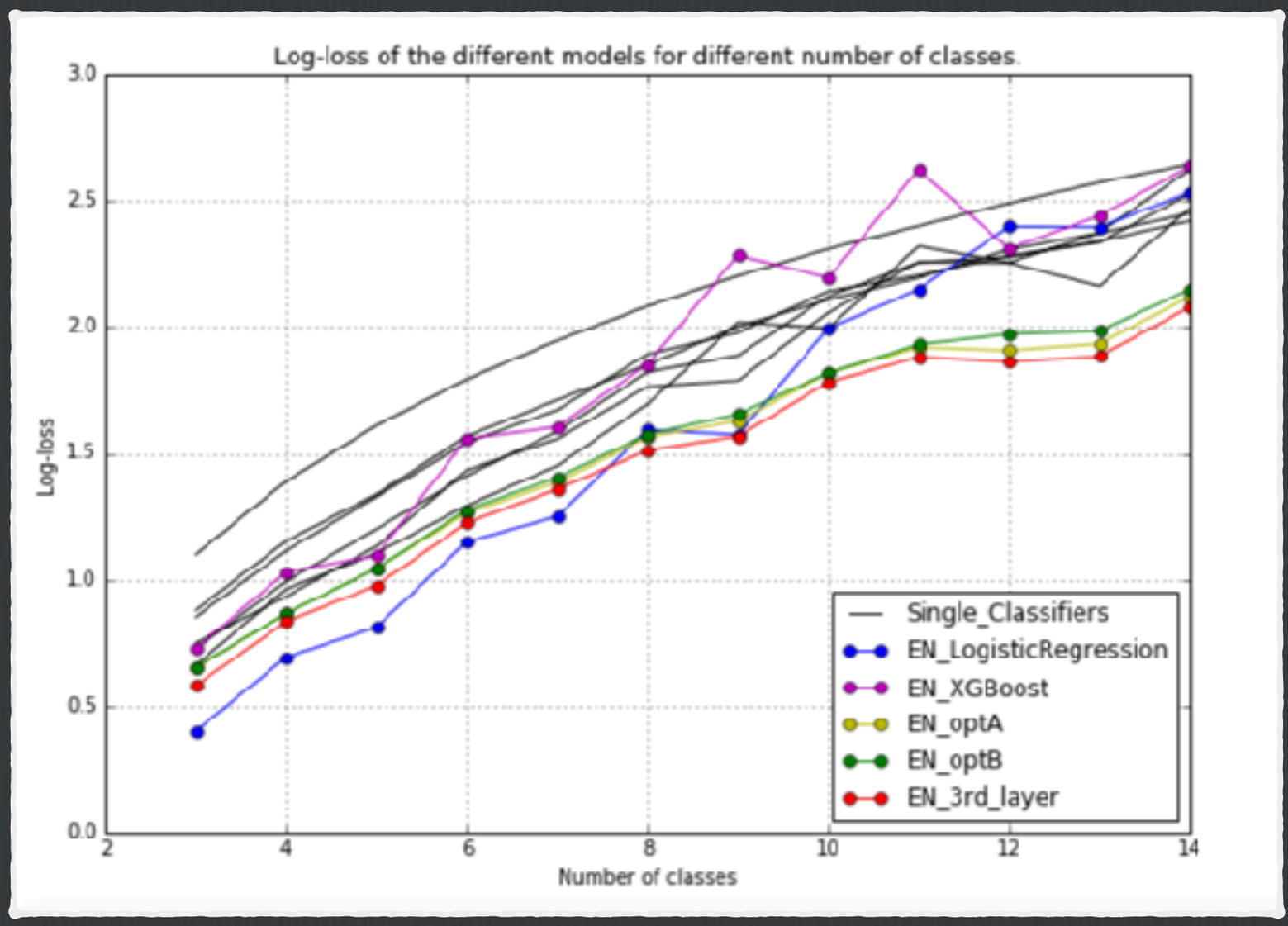
- Kaggle Challenge from 2015
- Goal: Predict the first travel destination of new US-based user
- Basic profile info, web session meta-data, additional facts and stats about demographics and countries

Training Data & Class Imbalance





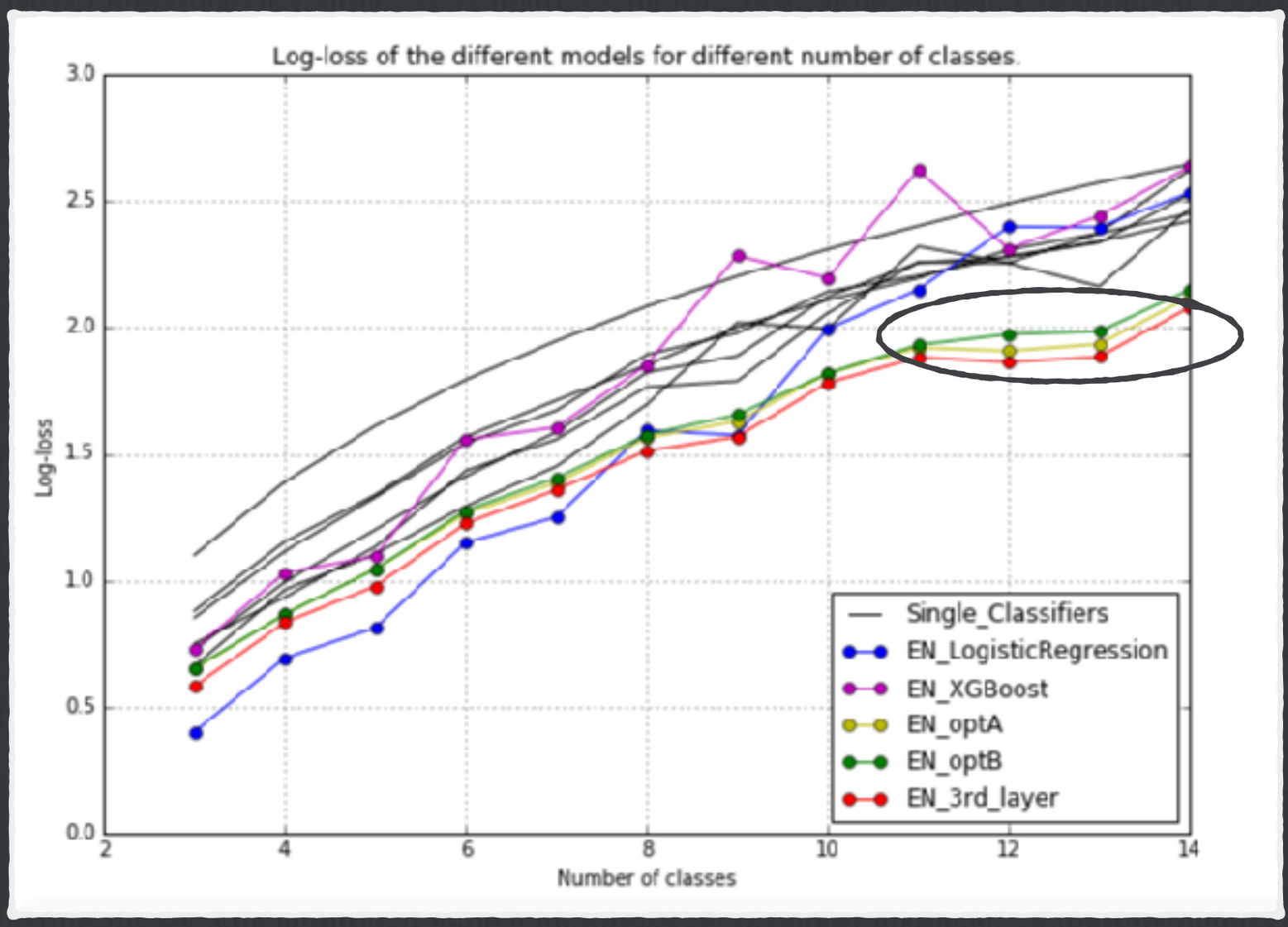
Previous Work & Metric



Sandro

- ☐ Key Metric is Log-Loss
- ☐ Lower Log-Loss is better
- ☐ Log-loss near 2.0 won

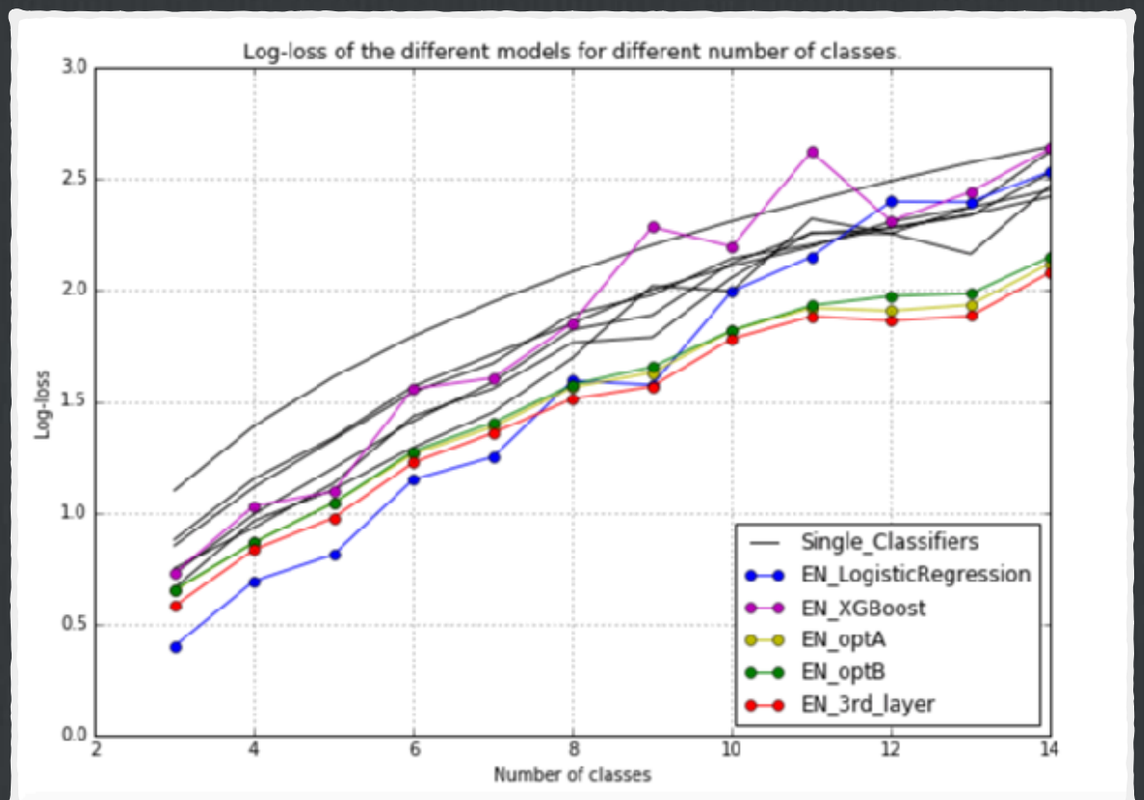
Previous Work & Metric



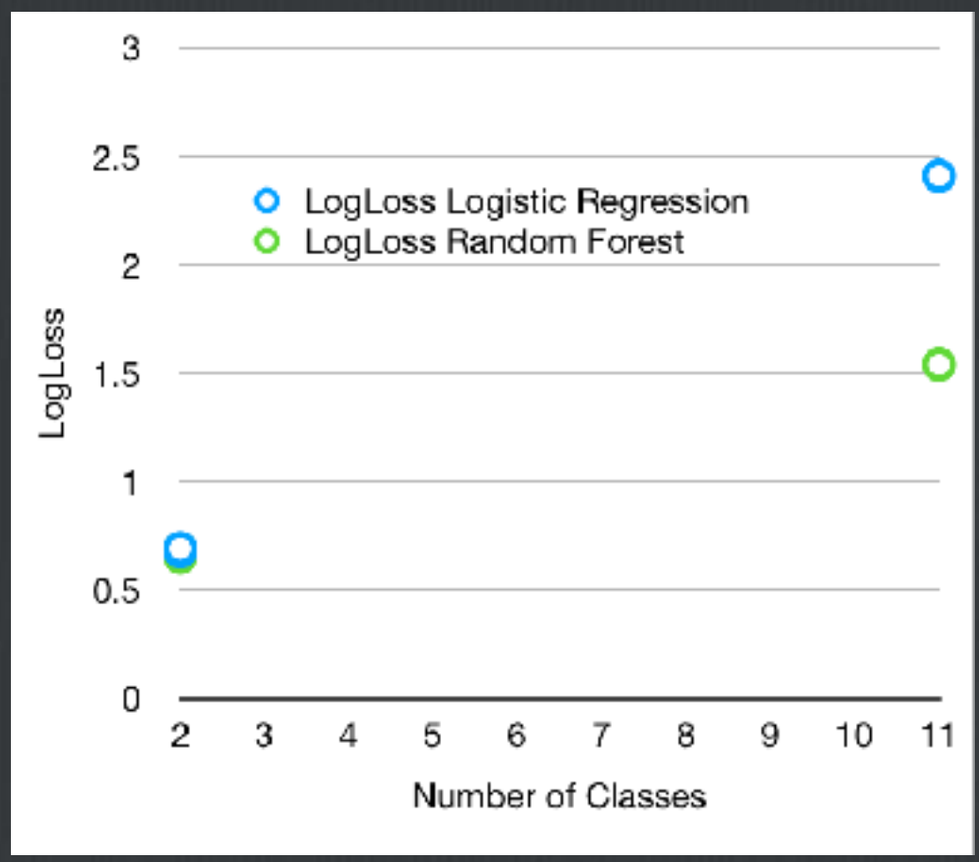
- ☐ Key Metric is Log-Loss
- ☐ Lower Log-Loss is better
- ☐ Log-loss near 2.0 won

Previous Work vs My Results

Kaggle 3rd place

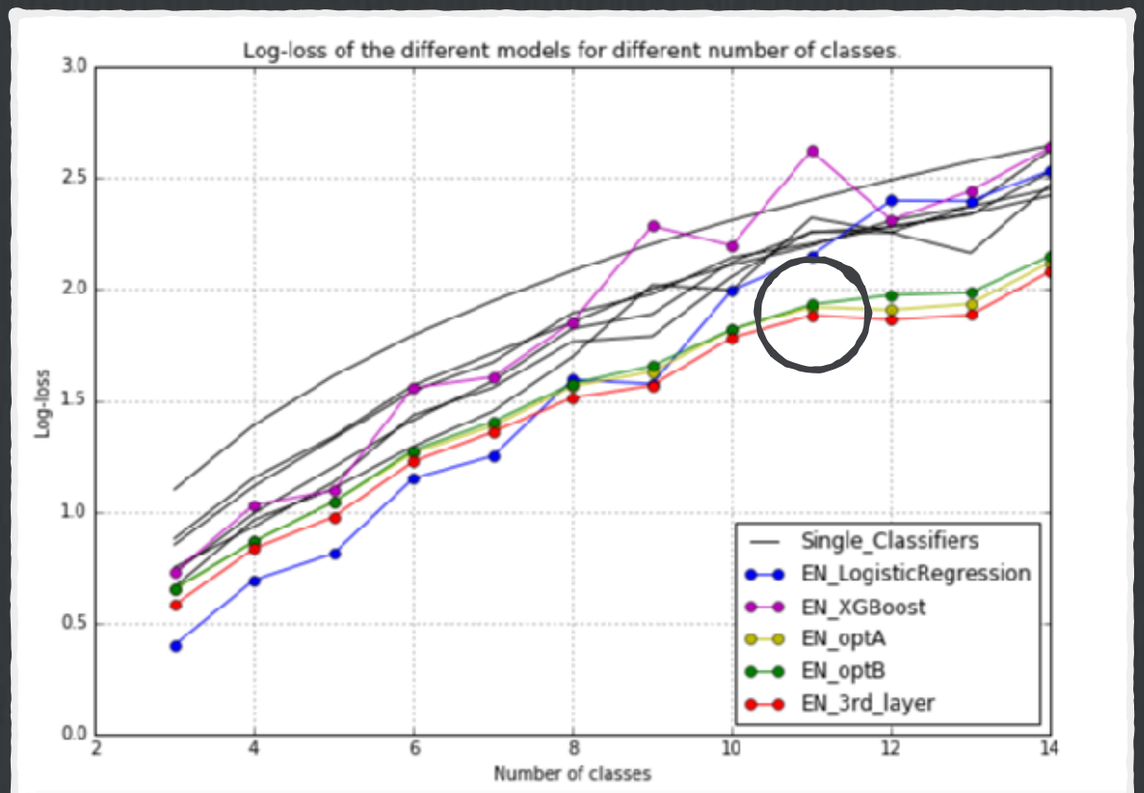


My results

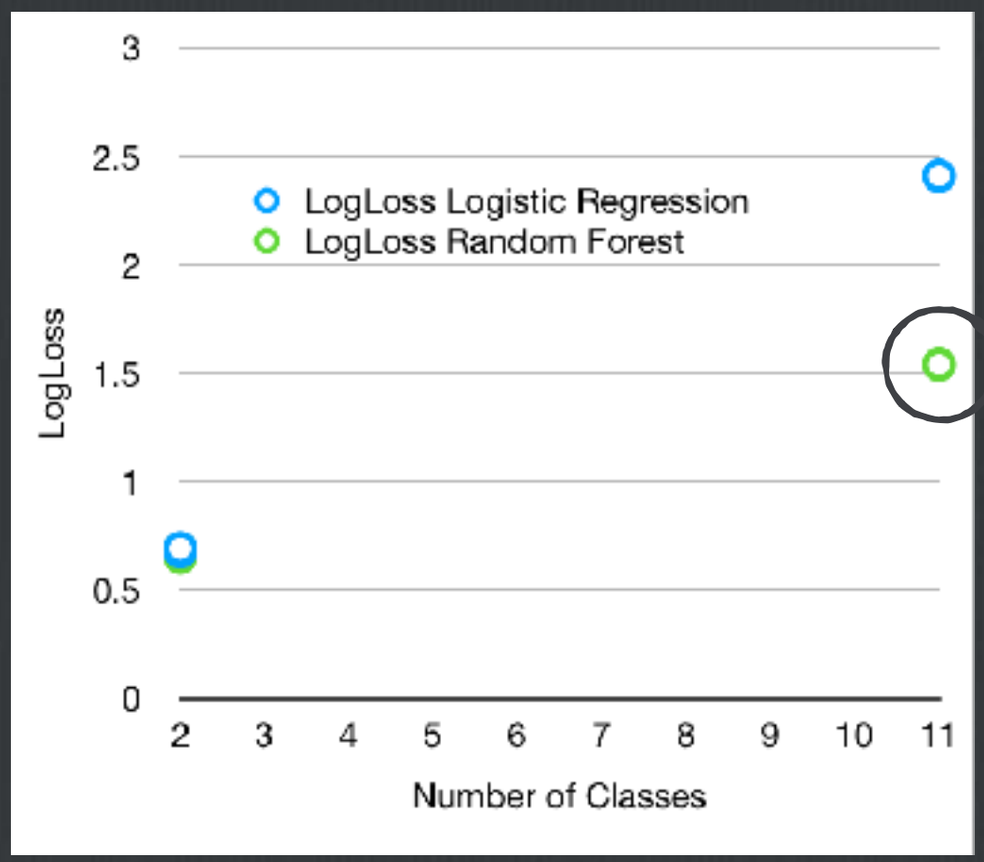


Previous Work vs My Results

Kaggle 3rd place



My results



Results: Major class USA

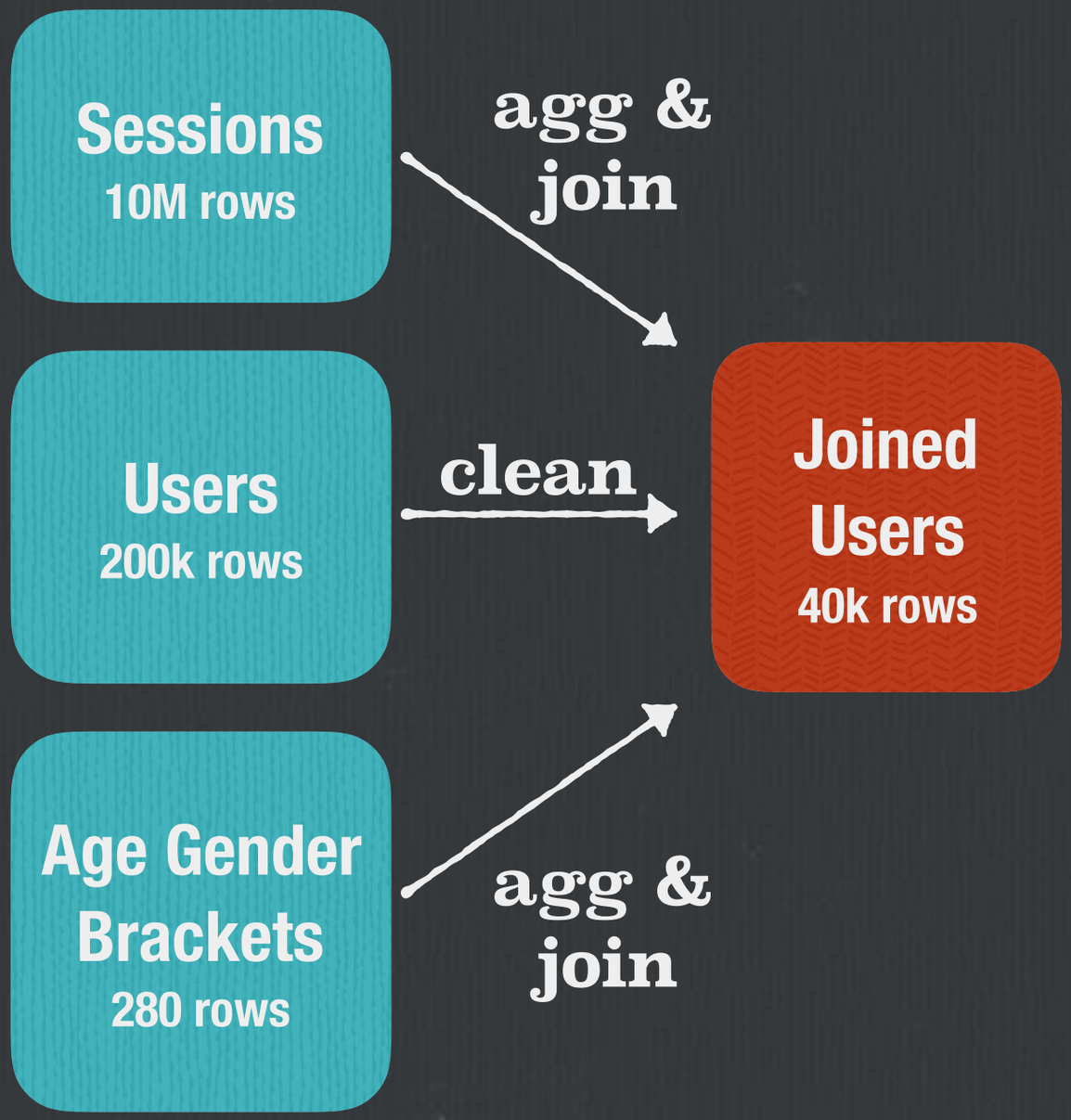
Country	precision	recall	f1-score	support
0	0.03	0.01	0.02	87
1	0.02	0.01	0.01	162
2	0.02	0.01	0.01	131
3	0.04	0.02	0.02	279
4	0.09	0.04	0.05	584
5	0.04	0.01	0.02	313
6	0.02	0.01	0.01	344
7	0.00	0.00	0.00	104
8	0.00	0.00	0.00	28
USA 9	0.71	0.90	0.79	7661
10	0.12	0.04	0.06	1209
avg / total	0.52	0.64	0.57	10902

Results: Minor Class France

Country	precision	recall	f1-score	support
0	0.03	0.01	0.02	87
1	0.02	0.01	0.01	162
2	0.02	0.01	0.01	131
3	0.04	0.02	0.02	279
4	0.09	0.04	0.05	584
5	0.04	0.01	0.02	313
6	0.02	0.01	0.01	344
7	0.00	0.00	0.00	104
8	0.00	0.00	0.00	28
9	0.71	0.90	0.79	7661
10	0.12	0.04	0.06	1209
avg / total	0.52	0.64	0.57	10902

France

ETL and Feature Engineering



odel
lection

Joined_Users

Train Val Holdout Split → Holdout

Train

Validation

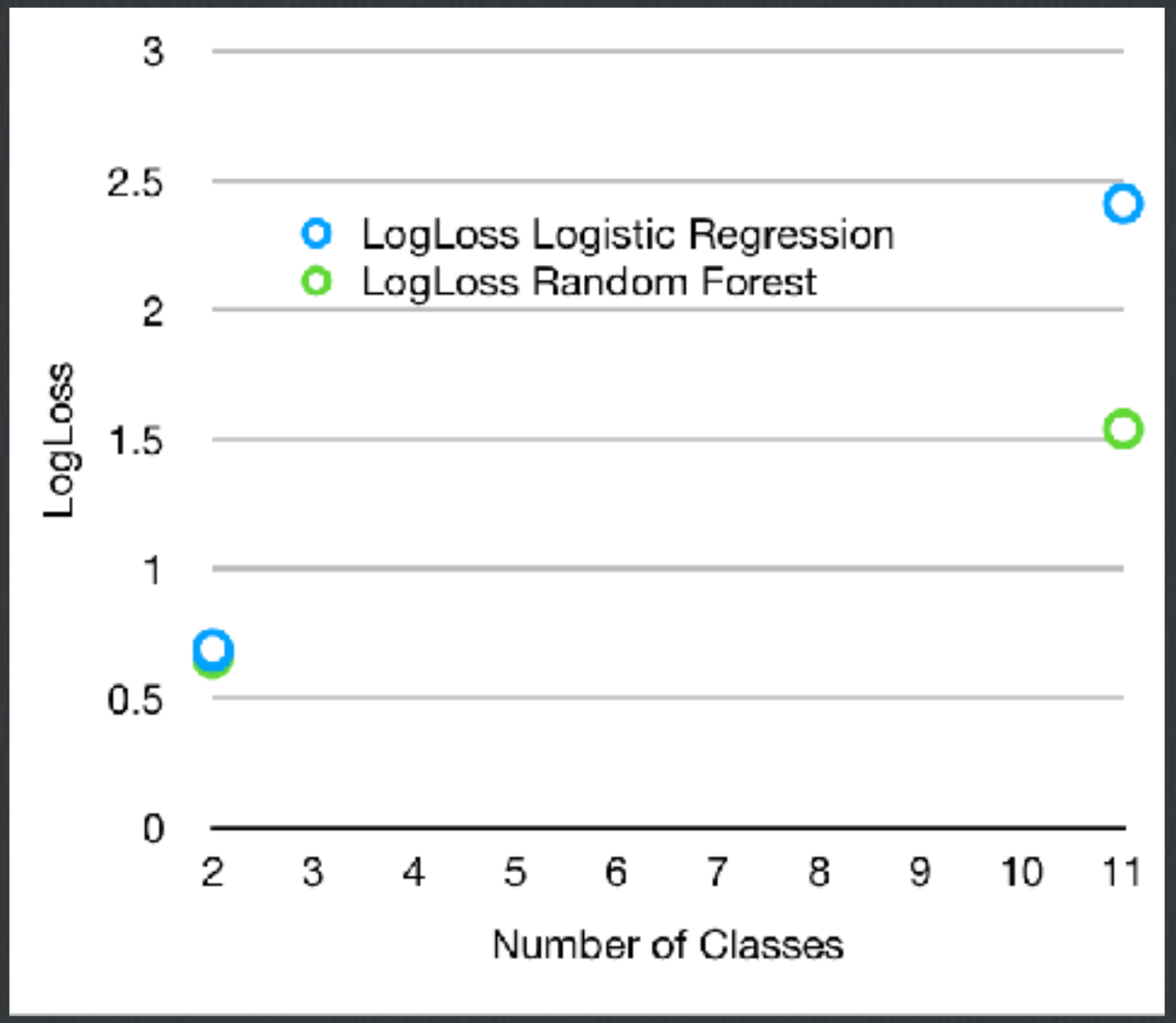
Transform, Binarized or Multilabel Subset

USA vs France
USA vs rest
Multilabel

SMOTE

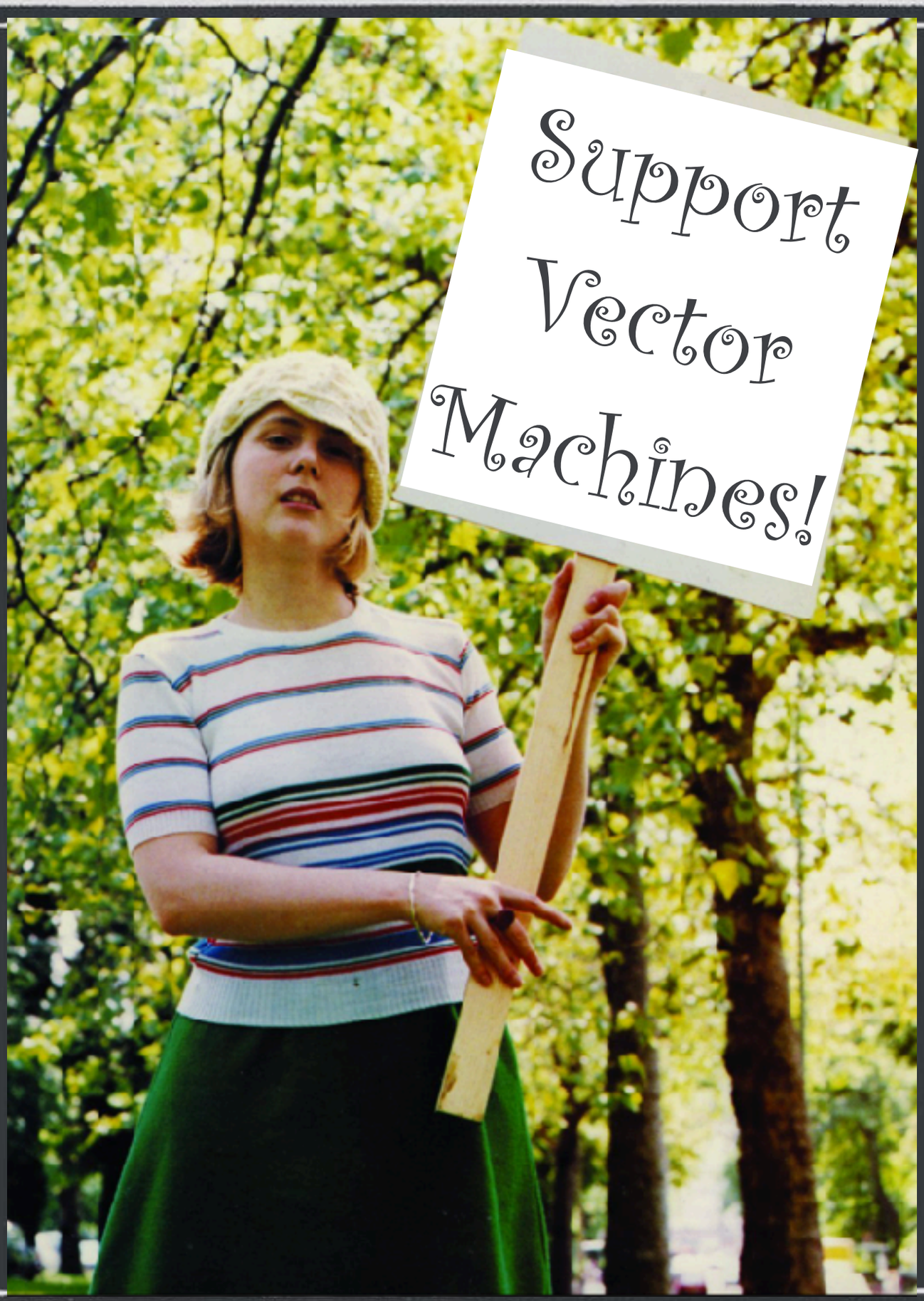
RandomizedSearchCV
- LogisticRegression
- RandomForest

Previous Work vs My Results



Conclusions and next steps

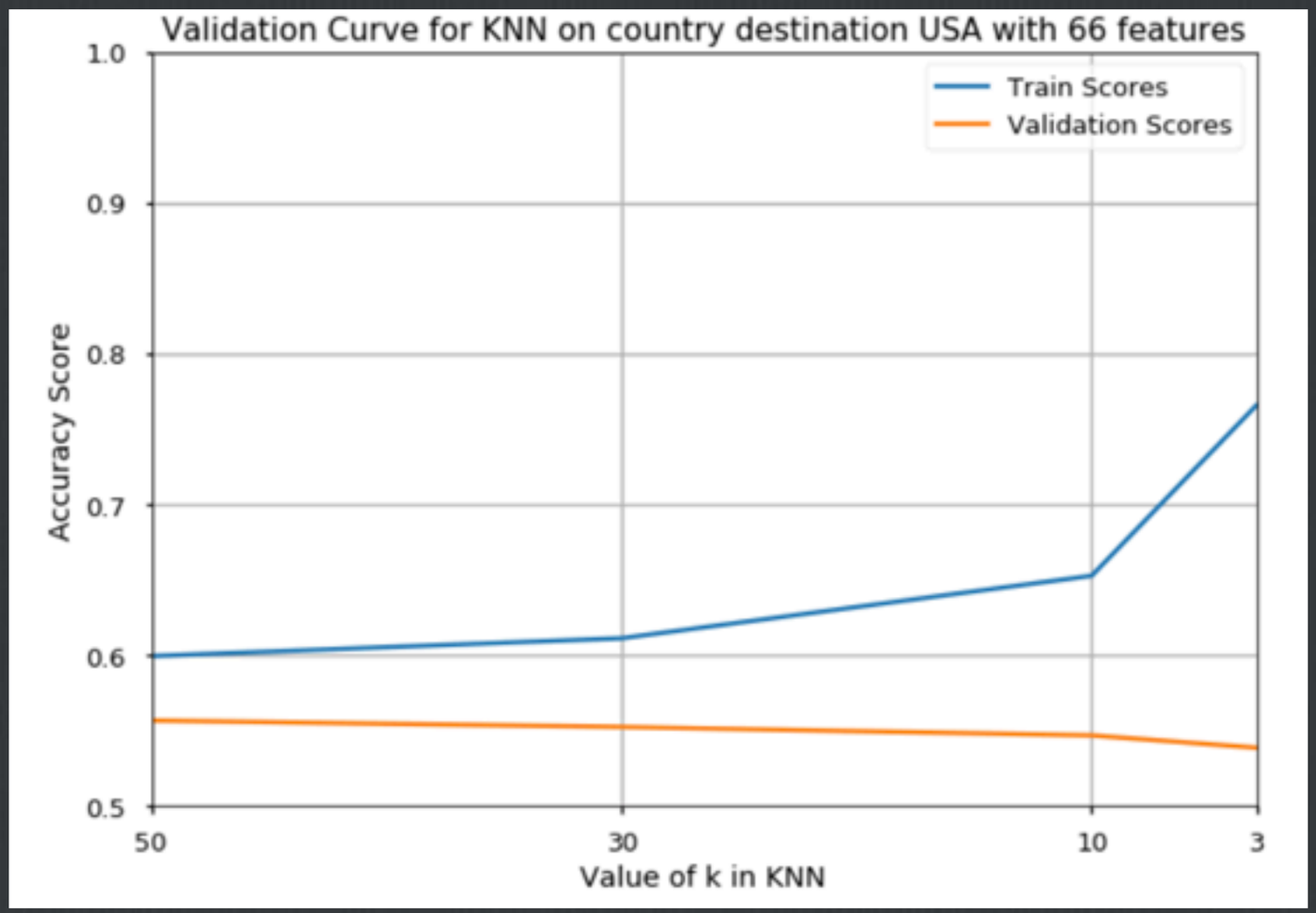
- ☐ AirBnB Data set provides low-grade information
- ☐ Ballpark of Kaggle 3rd Place Performance with simplified data set
- ☐ Next:
  - ☐ More models (XGBoost, SVM)
  - ☐ More data complexity
  - ☐ Understand Ensembling


Support
Vector
Machines!

Results: Major class USA

Country	precision	recall	f1-score	support
0	0.03	0.01	0.02	87
1	0.02	0.01	0.01	162
2	0.02	0.01	0.01	131
3	0.04	0.02	0.02	279
4	0.09	0.04	0.05	584
5	0.04	0.01	0.02	313
6	0.02	0.01	0.01	344
7	0.00	0.00	0.00	104
8	0.00	0.00	0.00	28
USA 9	0.71	0.90	0.79	7661
10	0.12	0.04	0.06	1209
avg / total	0.52	0.64	0.57	10902

Other stuff


Validation Curve for KNN on country destination USA with 66 features
Train Scores
Validation Scores
Accuracy Score
1.0
0.9
0.8
0.7
0.6
0.5
50
30
10
3
Value of k in KNN