MSE | MASTER OF SCIENCE IN ENGINEERING

BDA Project AirBNB

Salma CHERIF
Anthony HESSAB
Kim Leng CHHUN
Vincent ROBOTEL

09.06.2017

Plan

1. Context
2. Data
3. Data preprocessing
4. Applied models
5. Results
6. Conclusion
7. Further enhancements

1

Context

34 000+
cities

190+
country

Challenge

in which country a new AirBnB user will make his or her first booking?

- Share personalized content with community
- Decrease average time to first booking
- Better forecast demand

2

Data

- Source : Kaggle
- Volume : 200 000+ entries
- Format csv

- list of users, their demographics, web session records, and some summary statistics.

- All from the US

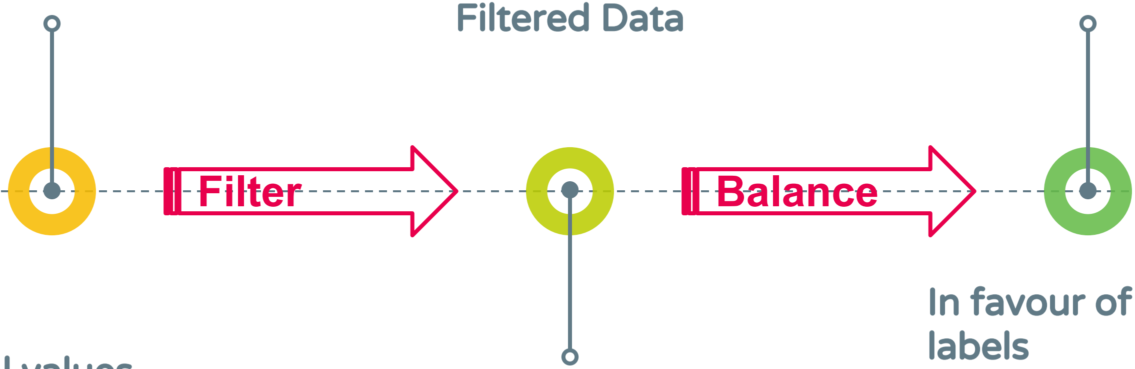
- 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF’, ‘other’

- Features
  - Gender
  - Age
  - Language
  - Account
  - Signup date
  - Browser
  - Device type
  - Affiliate channel …

3

Data preprocessing

Data preprocessing



Initial data

Null

Illogical values

Wrong / missing values

12 labels

213 451 entries

Filter

Filtered Data

Filtered/cleaned data
Normalization
Transformation
51 530 entries

Balance

Balanced Data

In favour of rare labels

10 230 entry

Label data distribution

Country destination	count
NL	460
PT	126
AU	349
CA	826
GB	1370
other	5660
DE	675
ES	1329
US	36314
FR	2868
IT	1553

After Filtering

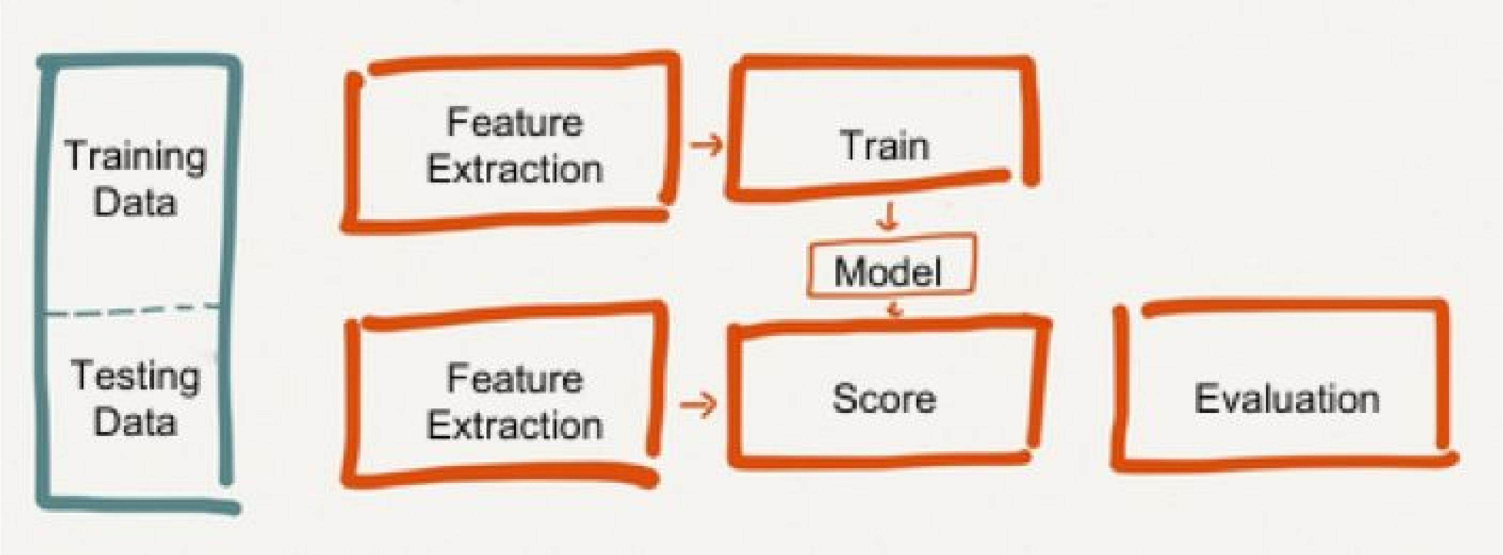
Country destination	count
NL	460
PT	126
AU	349
CA	826
GB	1299
other	1299
DE	675
ES	1299
US	1299
FR	1299
IT	1299

After Balancing

4

Applied models

Modeling process



Used models in chronology and difficulty

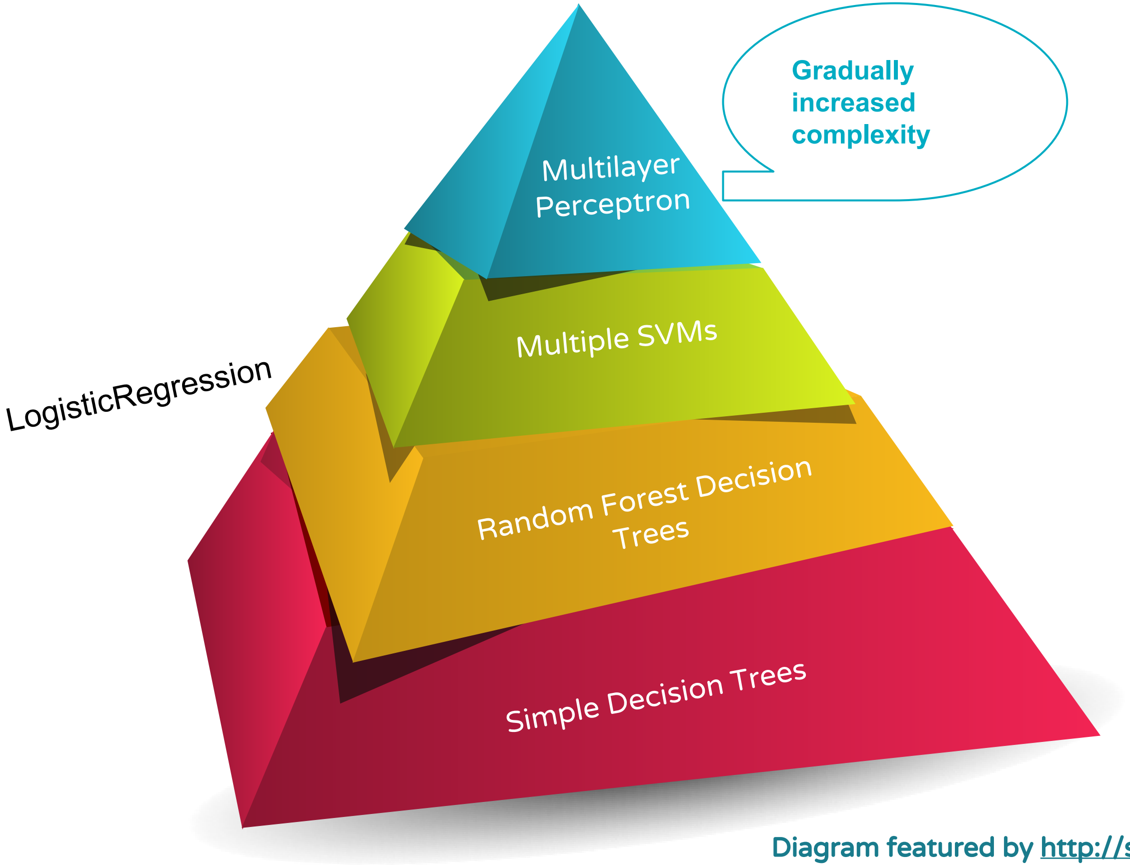


Diagram featured by http://slidemodel.com

5

Results

Accuracy

	Accuracy	Configuration
Decision Tree	0.7065614777756398	Impurity = Gini impurity, maxDepth=4, maxBins=100
Regression Logistic	0.7133436772692009	LogisticRegressionWithLBFGS
Random Forest	0.7052389176741508	Impurity = Gini impurity, maxDepth=24, maxBins=32
SVM	0.709056061	SVMWithSGD
Multilayer Perceptron	0.7091165715018926	.setLayers(Array(numColTrain-1, 25, 25, 20, dataValues.last.length)) .setBlockSize(256)
Multilayer Perceptron (more complex)	0.7053112139917695	.setLayers(Array(numColTrain-1, 256,256, 128,64, 64,32, dataValues.last.length)), .setBlockSize(100000)
Decision Tree (resampling)	0.1499032882011605З	Impurity = Gini impurity, maxDepth=4, maxBins=100

6

Conclusion

Conclusion

- Data “quality” was a handicap
- More features ?
- More data ?

- Applied several models with several configurations each
- Accuracy stuck at 0.7

7

Further enhancements

Further possible enhancements

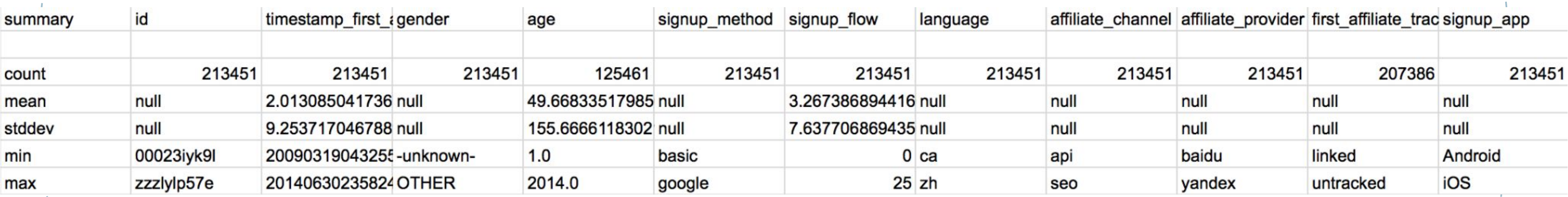
- Deal with missing values without truncating dataset (e.g. binning)
- Feature engineering
- Neural network [√]
  - Test done: 10 min and 4 hours training
  - More neurons and iterations

Thank you!

Any questions?

Annexes

Data profiling (original set)



summary	id	timestamp_first_a	gender	age	signup_method	signup_flow	language	affiliate_channel	affiliate_provider	first_affiliate_trac	signup_app
count	213451	213451	213451	125461	213451	213451	213451	213451	213451	207386	213451
mean	null	2.013085041736	null	49.66833517985	null	3.267386894416	null	null	null	null	null
stddev	null	9.253717046788	null	155.6666118302	null	7.637706869435	null	null	null	null	null
min	00023iyk9l	20090319043255	-unknown-	1.0	basic	0	ca	api	baidu	linked	Android
max	zzzlylp57e	20140630235824	OTHER	2014.0	google	25	zh	seo	yandex	untracked	iOS