amadeus

Relative Label Encoding for the Prediction of Airline Passenger Nationality

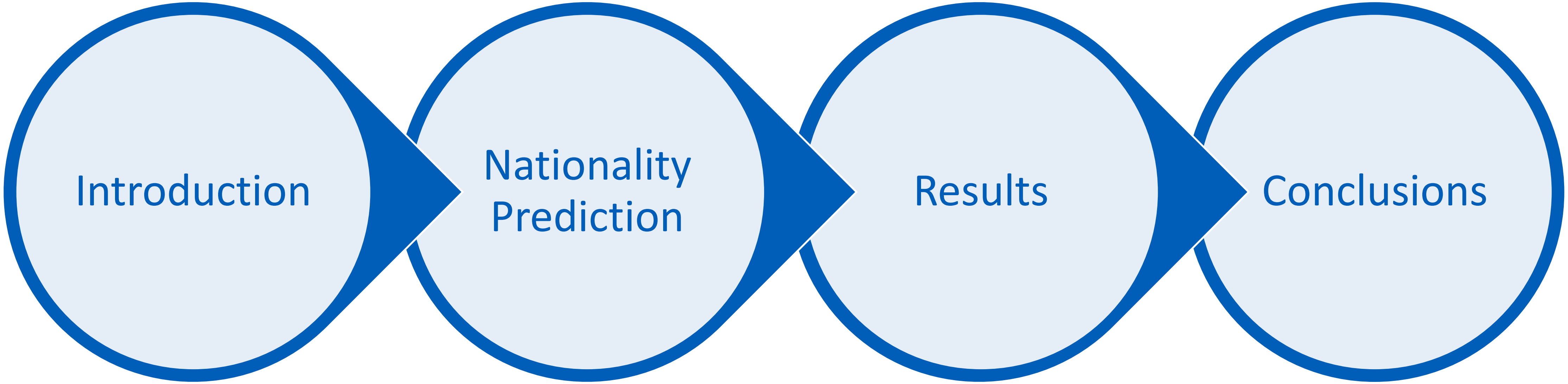
ICDM 2016
IEEE International Conference on Data Mining

DSBDA 2016
December 12th , 2016
Barcelona, Spain

Authors:
Alejandro MOTTINI, Ph.D.
Rodrigo ACUNA-AGOST, Ph.D.

Innovation & Research
Amadeus IT Group

Outline


Introduction
Nationality Prediction
Results
Conclusions

Outline

Introduction → Nationality Prediction → Results → Conclusions

Introduction

PNR Data

- PNR (Passenger Name Record):
  - Created when a travel reservation is made
  - Generated by airlines or authorized agents (i.e: travel agencies)
  - Once created, stored by the airlines and/or the Global Distribution Systems (like Amadeus)

- PNR contains:
  - **Travel itinerary (always present)**
  - Personal information (name, gender, age, etc)
  - Payment information (currency, total price, etc)
  - Other information (ancillary services, hotel reservation, etc)

```
RP/MUC1A0701/MUC1A0701            AA/CX  28NOV00/1419Z   KJ2LOI
  1.LAST_NAME1/NAME1(CHD)   2.LAST_NAME2/NAME2 (ADT)
  3  BA2817 C 23JUL 1 CDGLHR HK2       1  0700 0705   *1A/E*
  4  BA2716 C 03JUL 2 LHRFRA HK2       N  1340 1635   *1A/E*
  5  LH 738 C 03JUL 2 FRAHKG HK2  1720 1  1750 1050+1 *1A/E*
  6 ATX BA HN1 LHRLHR 04JUL-HELICOPTER TRANSFER/P1
  7 CCR ZE HK1 LHR 12JUL 13JUL ECMN/BS-00000000/ARR-0750
    /BN-12EF3436H/FT-111111111/ID-111111/RQ-PP/RT-1900/CF-/P1
    ** SEE RTSVCC**
  8 TUR BA HN1 LHR 16JUL-20JUL/VISIT SCOTLAND/P1
  9 HTL BA HN1 ZXE 21JUL-21JUL/EDINBURGH/P1
 10 IBOPEN C 01OCT 1 MADSCL
 11 AP LON 44 1 234 5678 H
 12 APS NCE0492940005-B
 13 TK OK28NOV/MUC1A0701
 14 SSR NSST BA HN2 CDGLHR/S3
 15 SSR NSST BA HN2 LHRFRA/S4
 16 SSR VGML BA HN2/S3
 17 *SSR FQTV BA HK/ IB00300004 SAPPHIRE/P2
 18 OSI YY 1CHD/P1
 19 OP MUC1A0701/28NOV/TEXT
 20 AI ANDEC00129
 21 RC MUC1A0701/THIS IS A CONFIDENTIAL REMARK
22 RM THIS IS A GENERAL REMARK
 23 RMC REQUEST NON SMOKING
 24 RMH REQUEST NON SMOKING
 25 RIA FRF2000-F
 26 RIS DEM25-SERVICE FEE
 27 RIT DEM340
 28 RII INVOICE AND ITINERARY
 29 RIR ITINERARY
 30 RIF INVOICE
 31 RQ THIS IS A QUALITY CONTROL REMARK
 32 FD AD50
 33 FE *M*PAY DIRECT TO VINCENT
 34 FM 9
 35 FP CASH
 36 FS 123AC
 37 FZ TICKET PAID BY IBM
 38 AB CY-GREAT COMPANY/NA-MR SMITH/A1-12 LONG STREET/ZP-BS7890/
       CI-NEWTOWN/CO-UNITED STATES/P2
 39 AM CY-MRMARTINEZDEMATA/1 SHORT STREET/ZP-BS7872/CI-NEWTOWN/CO-UNITED STATES/P2
```

Figure: Example of a PNR (illustrative example with fictitious data)

Introduction

PNR Data

- Passengers’ attributes are important to the industry: airports, airlines, travel agents
- Routinely used for different business applications:
  - Customer segmentation
  - Personalized product pricing
  - Adaptive airport personnel
- Problem: PNR data is not as complete as we would like …

	% of presence in PNR
Nationality	~10%
Age	~10%
Gender	~80%

Outline

Introduction

Nationality Prediction

Results

Conclusions

Nationality Prediction

Introduction

- Complicated problem given the information present in PNR
  - Itinerary
  - Currency
  - Country office id

- **Challenges:**
  - 195 classes
  - Unbalanced data (3 countries make up 57% of records)
  - Nationalities distribution varies significantly from airport to airport
  - Some cases are not predictable

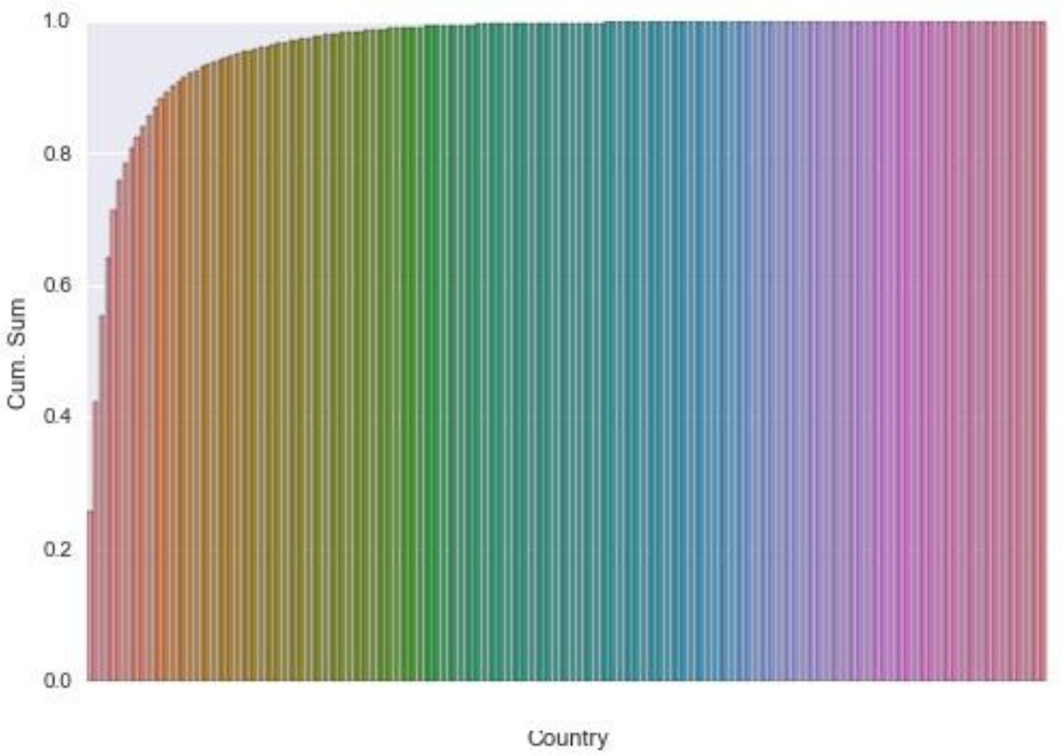


Figure: Cumulative sum of the number of records for the different nationalities present in the considered dataset. Countries have been anonymized

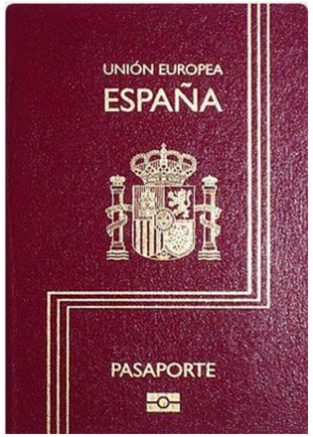
Nationality Prediction

Introduction

- Real example:
  - Outbound Trip: NCE **(FR)** -> BCN **(ES)**
  - Return trip (5 days later): BCN **(ES)** -> NCE **(FR)**
  - Currency: **EUR**
  - Office Id country: **FR**

- Guessed Nationality:
  - French ?
  - Spanish ?
  - Other?

?



Nationality Prediction

Introduction

- Real example:
  - Outbound Trip: NCE **(FR)** -> BCN **(ES)**
  - Return trip (5 days later): BCN **(ES)** -> NCE **(FR)**
  - Currency: **EUR**
  - Office Id country: **FR**

- **Answer**: IT citizen

Nationality Prediction

Alternative methods

- Nationality is usually predicted using:
  - Ad-hoc rule-based methods
  - Estimated using surveys

- **Shortcomings**:
  - **Rule-based methods**:
    - Lower accuracy than ML models
    - Impractical to optimize (nationality distribution varies from airport to airport)
  - **Surveys**:
    - Expensive
    - Time consuming
    - Not reactive (analysts receive information one month after travel date)

Relative Label Encoding

Nationality prediction

- Take advantage of the fact that class labels (country codes) are represented in the same space as the features used to predict it

- Passengers are assigned to four classes
  - Nationality == Country Origin Trip (**class 0**)
  - Nationality == Country Destination Trip (**class 1**)
  - Nationality == Currency of purchase (**class 2**)
  - Nationality == None of the above (**class 3**)

- Relative label encoding transforms the original high cardinality label space into one with 4 labels

- New classes are non-exclusive -> **Multi-label Classification**

Relative Label Encoding

Encoding Example

Origin	Destination	Currency	Others	Nationality
US	CN	USD	…	US
FR	DE	EUR	…	DE
ES	IT	EUR	…	NL
MX	CL	USD	…	AR
US	US	USD	…	US

Data in the original label space. "Others" column represent additional features that are used for the prediction but are not relevant for the encoding

Encoding →

Origin	Destination	Currency	Others	Nationality
US	CN	USD	…	(0,2)
FR	DE	EUR	…	(1,2)
ES	IT	EUR	…	(2)
MX	CL	USD	…	(3)
US	US	USD	…	(0,1,2)

Target variable after encoding. Labels are the index of the feature it matches

Relative Label Encoding

Multi-label Classification

- Multi-label Classification: Multi-class classification where an instance can belong to many non exclusive classes

- Two main approaches:
  - Algorithm adaptation: adapting an existing single-label algorithm
  - Problem transformation: transform original problem into several binary classification ones

- Binary relevance (BR):
  - One binary classifier is independently trained for each label.
  - Efficient and easy to implement
  - Shortcoming: Assumes label independence

Relative Label Encoding

Multi-label Classification

- Classifier Chain (CC):
  - As in BR, CC uses L binary classifiers
  - Binary classifiers are linked along a chain
  - Feature space of each link (binary classifier) is incremented with the 0/1 label associations of previous links
  - In prediction time, classification of a new instance is carried out as in BR, but using the augmented binary classifiers
  - **Shortcoming**: result depends on chose order of chain

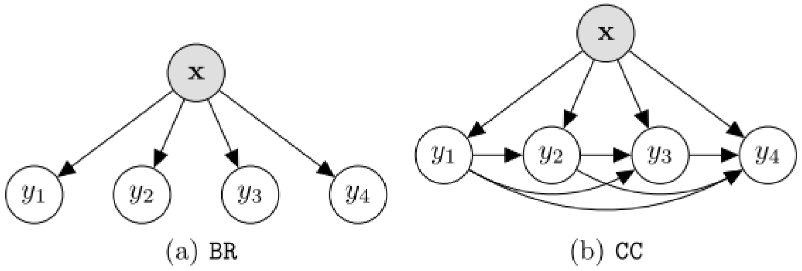


Fig. 1: BR (1a) and CC (1b) as graphical models, $L = 4$.

From: A Deep Interpretation of Classifier Chains, Read and Hollmen, 2014.

- CC can me combined into an ensemble (ECC):
  - ECC trains several CC classifiers, each one using a different random chain order
  - Increases overall accuracy and reduces over-fitting.
  - Compromise between computational complexity and capturing label correlation

amadeus

Relative Label Encoding

Label Decoding

- Multi-label classification algorithm for prediction
- Predicted (encoded) labels need to be transformed back into the original label space (country code)

- Decoding procedure:
  - For each passenger, trained model predicts the probability of each of the 4 encoded labels
  - Each encoded label is transformed back into the original country code
  - Class probabilities are used as weights for the countries
  - Country codes are grouped using the sum of the weights
  - The country with the biggest weight is chosen as the final predicted nationality

- In case of a tie:
  - Country is randomly chosen from the ones sharing the highest total weight

- If class 3 (nationality = country XX) wins:
  - Most frequent nationality excluding those matching the countries of origin/destination/currency is chosen.

Relative Label Encoding

Label Decoding Example

Origin	US
Destination	CN
Currency	USD
Other Features	…
Prob. Class 0	*0.4*
Prob. Class 1	*0.2*
Prob. Class 2	*0.3*
Prob. Class 3	*0.1*

→

Origin	US
Destination	CN
Currency	USD
Other Features	…
Prob. Class 0	*0.4* ***US***
Prob. Class 1	*0.2* ***CN***
Prob. Class 2	*0.3* ***US***
Prob. Class 3	*0.1* ***XX***

Group by country (agg=sum)

→

Origin	US
Destination	CN
Currency	USD
Other Features	…
Predicted Nationality	**US**

Class 0: Nationality = Country Origin
Class 1: Nationality = Country Destination
Class 2: Nationality = Country Currency
Class 3: Nationality = Other (Country XX)

Outline

Introduction

Nationality Prediction

Results

Conclusions

Results

Results

- Validation on PNRs of passengers passing through an important European airport
- One month of data, only certain airlines (aprox. 100K records)
- Data preprocessing with Spark cluster, ML locally with Scikit-learn (for first prototype)
- Hyper-parameter optimization using grid search and 5-fold cross validation

Results

Evaluation Metrics

- **Overall accuracy**: $ACC = \frac{TP + TN}{P + N}$
- **Percentage of detected nationalities**: detected percentage (relative to all nationalities in training set)

- **Balanced error rate**:
  - Uniform average of the proportion of wrong classifications in each class (useful for skewed data)

	0	1	2
0	a	b	c
1	d	e	f
2	g	h	i

$$BER = \frac{\frac{(b+c)}{(a+b+c)} + \frac{(d+f)}{(d+e+f)} + \frac{(g+h)}{(g+h+i)}}{3}$$

- **Weighted average error per nationality**:
  - Main concern is to correctly approximate the distribution of nationalities and not to correctly predict the nationality of a particular passenger **(used by airport analysts)**

$$W_{AEN} = \frac{\sum_{i=1}^{N} w_i |w_i - pred_i|}{\sum_{i=1}^{N} w_i}$$

W_i = total passengers with real nationality *i*
Pred_i = total passengers with predicted nationality *i*

Results

Results

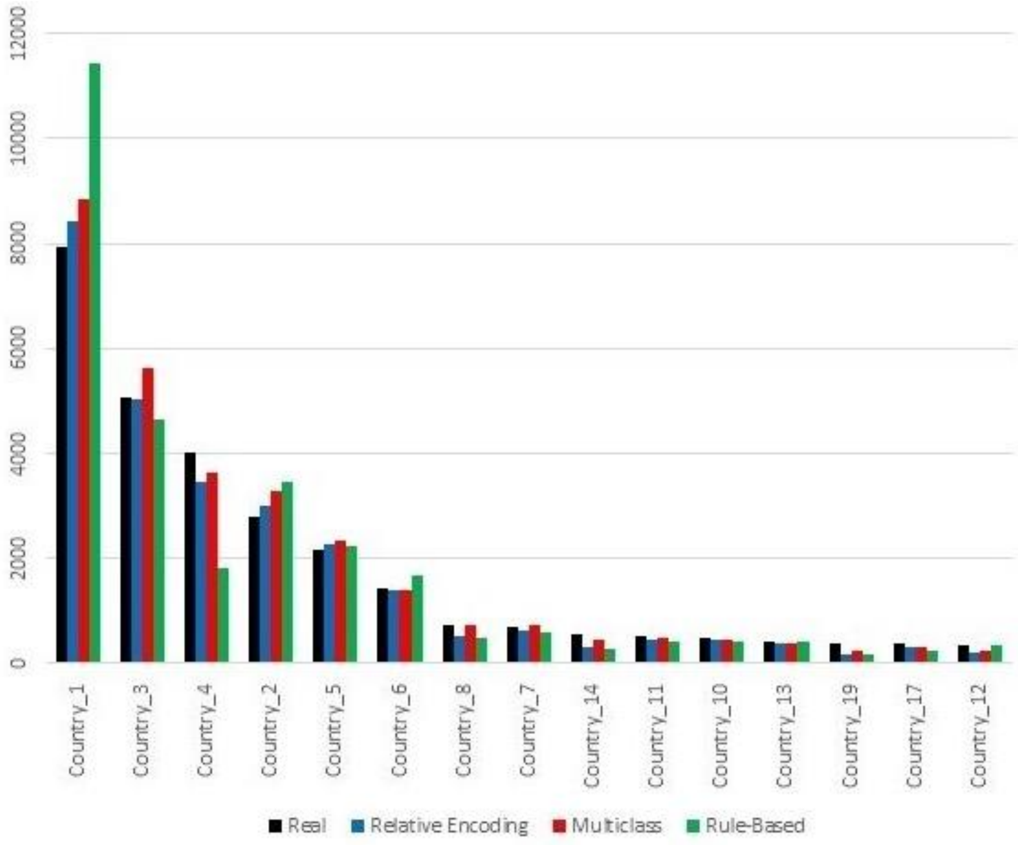
Method	Overall Acc.	BER	W_{AEN}	% Detected Nat.
Relative encoding	**0.78**	**0.38**	**0.13**	69%
Multi-class	0.77	0.46	0.14	56%
Rule based	0.68	0.38	0.31	**73%**

- Evaluation in original label space (country codes)
- Multi-class: Classical classification in the original country code space
- Rule-based: predicted nationality equals country of origin

- **Relative encoding obtains the best overall performance**

Results

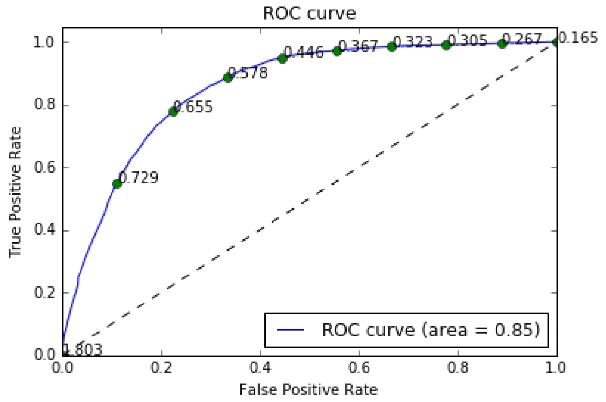
Results



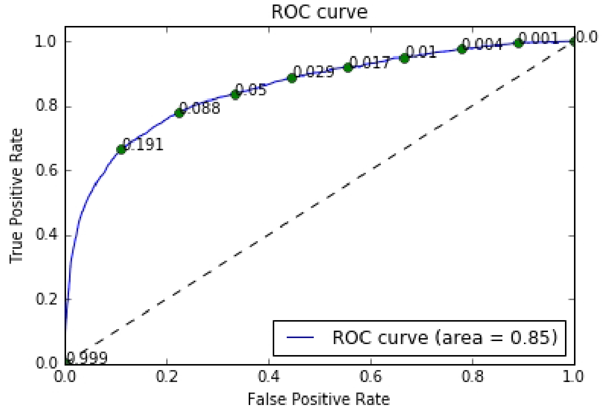
Predicted number of passengers per nationality for each method, comparison with ground truth (the countries have been anonymized and ordered by number of passengers)

Results

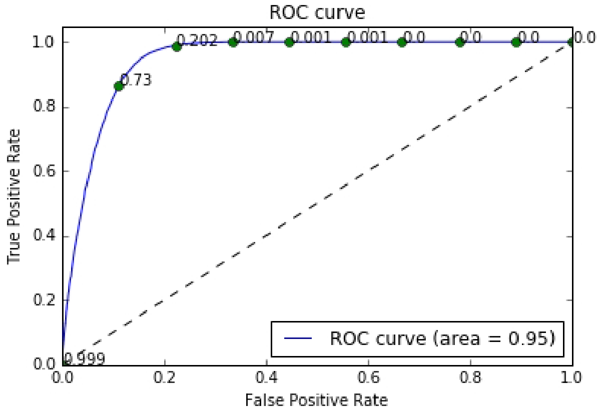
Results



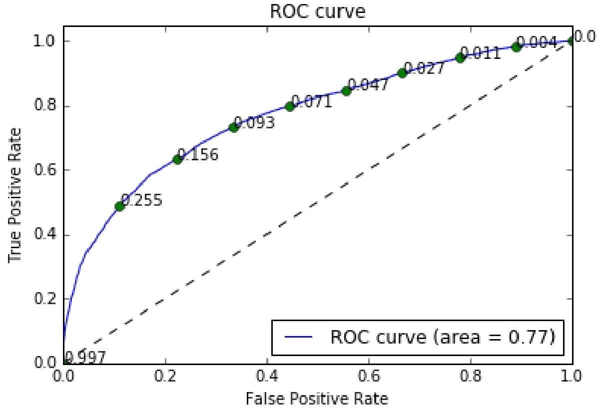
Class 0



Class 1



Class 2



Class 3

Outline

Introduction → Nationality Prediction → Results → **Conclusions**

Conclusions

Problem

- Airports/Airlines are interested in knowing passengers‘ nationality
- Used for different commercial applications
- Can be present in PNR data
- Only 10% of PNR have this information

What we propose

- Method to predict the nationality of passengers based on PNR data
- Take advantage of a particularity of this type of data
- Encode the target variable by assigning it the index of the feature it matches
- Passengers can belong to one of four classes (instead of 195)
- Non-exclusive classes -> Multi-label classification

Results

- Evaluated on a PNR dataset of travelers passing through an important European airport
- Proposed method outperforms simple multi-class approach and a rule-based method
- Relative encoding is more flexible than multi-class classification
- Able to predict nationalities unseen by the model

amadeus

Strictly CONFIDENTIAL

Thank you!

Classifier Chain

TRAINING$(D = \{(x_1, S_1), \cdots, (x_n, S_n)\})$

for $j \in 1 \cdots |L|$
  do $\triangleright$ single-label transformation and training
    $D' \leftarrow \{\}$
    for $(x, S) \in D$
      do $D' \leftarrow D' \cup ((x, l_1, \cdots, l_{j-1}), l_j)$
    $\triangleright$ train C_j to predict binary relevance of l_j
    $C_j : D' \rightarrow l_j \in \{0, 1\}$

Fig. 1: CC's training phase for dataset D and label set L.

CLASSIFY(x)

$Y \leftarrow \{\}$
for $j \leftarrow 1$ **to** $|L|$
  do $Y \leftarrow Y \cup (l_j \leftarrow C_j : (x, l_1, \cdots, l_{j-1}))$
return (x, Y) $\triangleright$ the classified example

Fig. 2: CC's prediction phase for a test instance x.

- BR: $O(|L| \times f(|X|, |D|))$
- CC: $O(|L| \times f(|X| + |L|, |D|))$

X → Features; Y → L1

Features	L1
...	1
...	0
...	0

C1: X-> Y

X → Features, L1; Y → L2

Features	L1	L2
...	1	0
...	0	1
...	0	0

C2: X-> Y

Ensemble of Classifier Chains

- ECC: trains m CC chains, each one with different order and subset of D
- For each data point, each CC model predicts: $y_k = (l_1, \cdots, l_{|L|}) \in \{0,1\}^{|L|}$
- The results are aggregated for all CC: $W = (\lambda_1, \cdots, \lambda_{|L|}) \in \mathbb{R}^{|L|}$, $\lambda_j = \sum_{k=1}^{m} l_j \in y_k$
- Each lambda contains the votes this label received
- A threshold t is used to choose the final multi-label set Y such that : $l_j \in Y$ where $\lambda_j \geq t$

Features

TABLE I
FEATURE NAMES AND TYPES USED FOR THE NATIONALITY PREDICTION.

Feature	Type
Country Origin Trip	Categorical
Country Destination Trip	Categorical
One Way Trip	Numerical (binary)
Currency Purchase	Categorical
Country Office Id	Categorical
Stay Saturday	Numerical (binary)
Purchase Anticipation	Numerical
Number Passengers	Numerical
Travelling With Children	Numerical (binary)
Country Origin == Country Destination	Numerical (binary)
Country Origin == Country Office Id	Numerical (binary)
Country Origin == Country Currency	Numerical (binary)
Country Destination == Country Office Id	Numerical (binary)
Country Destination == Country Currency	Numerical (binary)
Country Office Id == Country Currency	Numerical (binary)

References

- S. Chen, J. Zhu, Q. Xie, W. Huang and Y. Huang, Understanding Airline Passenger Behavior through PNR, SOW and Webtrends Data Analysis, In Proc. International Conference on Big Data Computing Service and Applications, 2015
- B. Vinod, The continuing evolution: Customer-centric revenue management, Journal of Revenue and Pricing Management, vol 7(1), pp 27-39, 2008
- International Civil Aviation Organization, *Doc 9944, Guidelines on Passenger Name Record (PNR) Data*, 2010
- Unison Consulting Inc., Los Angeles International Airport 2011 Passenger Survey, Results and Findings. Available online: https://www.lawa.org/uploadedFiles/OurLAX/pdf/LAX Survey Final/Draft REPORT 2012 08 19.pdf
- J. Read, B. Pfahringer, G. Holmes and E. Frank, *Classifier Chains for Multi-label Classification*, Machine Learning and Knowledge Discovery in Databases, vol 5782, pp 254-269, 2009
- J.Y. Jiang, S.C. Tsai and S.J Lee, *FSKNN: Multi-label text categorization based on fuzzy similarity and k nearest neighbors*, Expert Systems with Applications, vol 39(3), pp 28132821, 2012
- C. Vens, J. Struyf, L. Schietgat, S. Dzeroski, and H. Blockeel, *Decision trees for hierarchical multilabel classification*, Machine Learning, vol 73(2), pp 185214, 2008
- A. Clare and R.D. King, *Knowledge Discovery in Multi-Label Phenotype Data*, In Proc. 5th European Conference on Principles of Data Mining and Knowledge Discovery, 2001
- G. Tsoumakas and I. Katakis, *Multi label classification: an overview*, International Journal of Data Warehouse and Mining, vol 3, pp 113, 2007
- J. Friedman, *Greedy function approximation: a gradient boosting machine*, Annals of Statistics, vol 29(5), pp 11891232, 2001
- Y. Freund and R.E. Schapire, *A decision-theoretic generalization of online learning and an application to boosting*,Journal of computer and system sciences, vol 55(1), pp 119-139, 1997
- T. Chen and C. Guestrin. *XGBoost: A Scalable Tree Boosting System*. In Proc. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016.
- United Nations Rules for Electronic Data Interchange for Administration, Commerce and Transport, *Introducing UN/EDIFACT*, Online at http://www.unece.org/cefact/edifact/welcome.html
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel,B. Thirion, O. Grisel,M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher,M. Perrot and E. Duchesnay, *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, vol 12, pp 2825-2830, 2011
- X. Meng, J. K. Bradley, B. Yavuz, E. R. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. B. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar, *MLlib: Machine Learning in Apache Spark*, Journal of Machine Learning Research, 2016.
- I. Guyon, A.R.S. Azar Alamdari, G. Dror and J. Buhmann, *Performance Prediction Challenge*, In Proc. IEEE International Joint Conference on Neural Network, 2006

Introduction

Amadeus

- Amadeus is a technology company dedicated to the global **travel industry**
- We are present in **195** countries
- Worldwide, we are **14000**+ people
- Our solutions help improving the business performance of: travel agencies, corporations, airlines, airports, hotels, railways and more.