Resampling Strategies for Imbalanced Time Series

Nuno Moniz, Paula Branco, Luís Torgo

LIAAD – INESC Tec
Sciences College, University of Porto

3rd International Conference on Data Science and Advanced Analytics
(DSAA 2016)

Introduction

- Time Series: sequence of data points; successive measurements over a time interval

- Past and Future Values Correlation

- Mining time series data is one of the most challenging problems in the field of data mining.

- The objective in many forecasting tasks involving time series is predicting rare values.

Prediction of Rare Values

- A common issue is the imbalanced distribution of the target variable

- Standard learning algorithms bias the models towards the most frequent situations, away from user preference biases.

- Examples: Financial data analysis, intrusion detection in network forensics, oil spill detection, prognosis of machine failures, ...

Problem Definition

- Time series forecasting assumes the availability of a time-ordered set of observations of a given continuous variable, Y

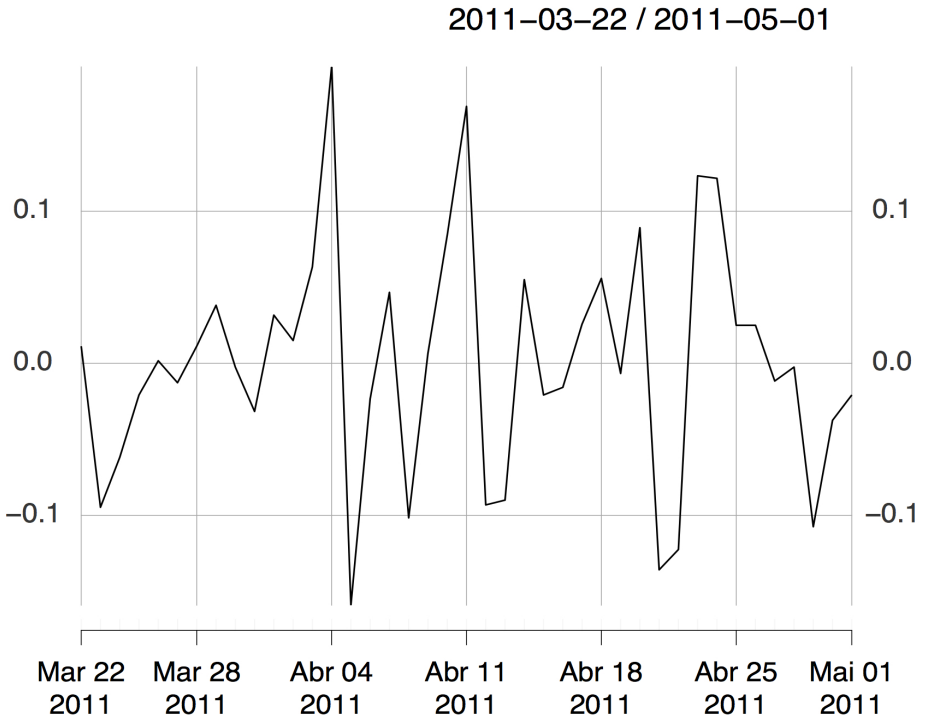
- The overall assumption is that an unknown function correlated past and future values of Y

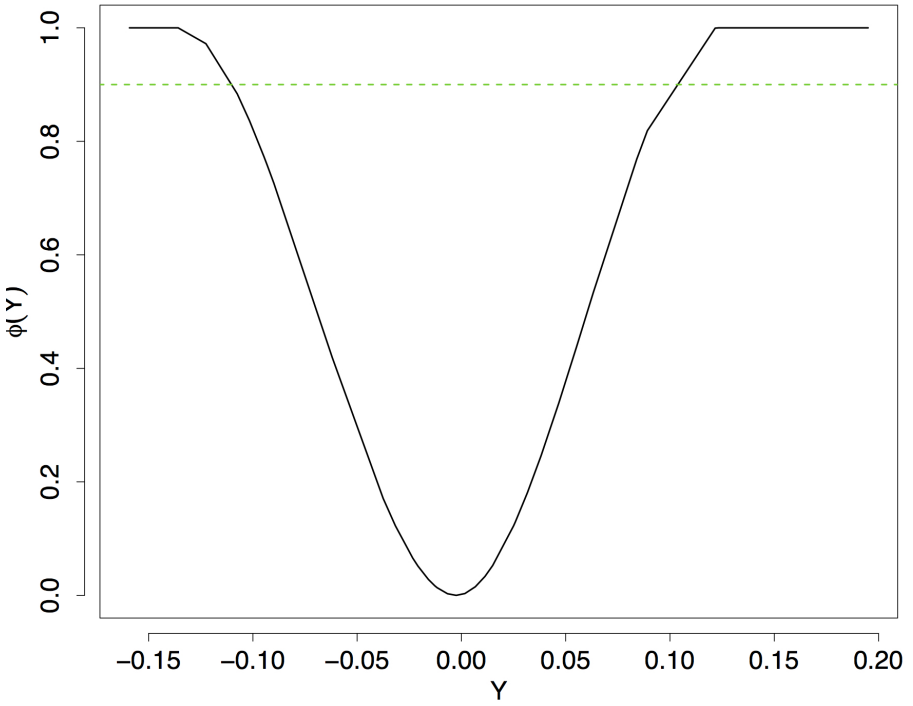
- A common form of modelling this correlation is time delay embedding [Takens, 1981]

- Unlike standard tasks, our predictive accuracy focus is towards the most relevant items

Relevance?

- Ribeiro [Ribeiro, 2011] proposes an approach to obtain a relevance function that maps the domain of continuous variables into a [0,1] scale of relevance: $\phi(Y)$: $Y \rightarrow [0,1]$





Resampling Strategies

- Pre-processing step

- Common in classification tasks, also extended to regression tasks [Torgo, 2007]

- Consists of removing normal (non-relevant) and/or adding rare (relevant) cases

Random Undersampling: balances the number of normal and rare values by randomly removing normal cases

SMOTE: combines random undersampling with oversampling through the generation of synthetic cases

Temporal Bias

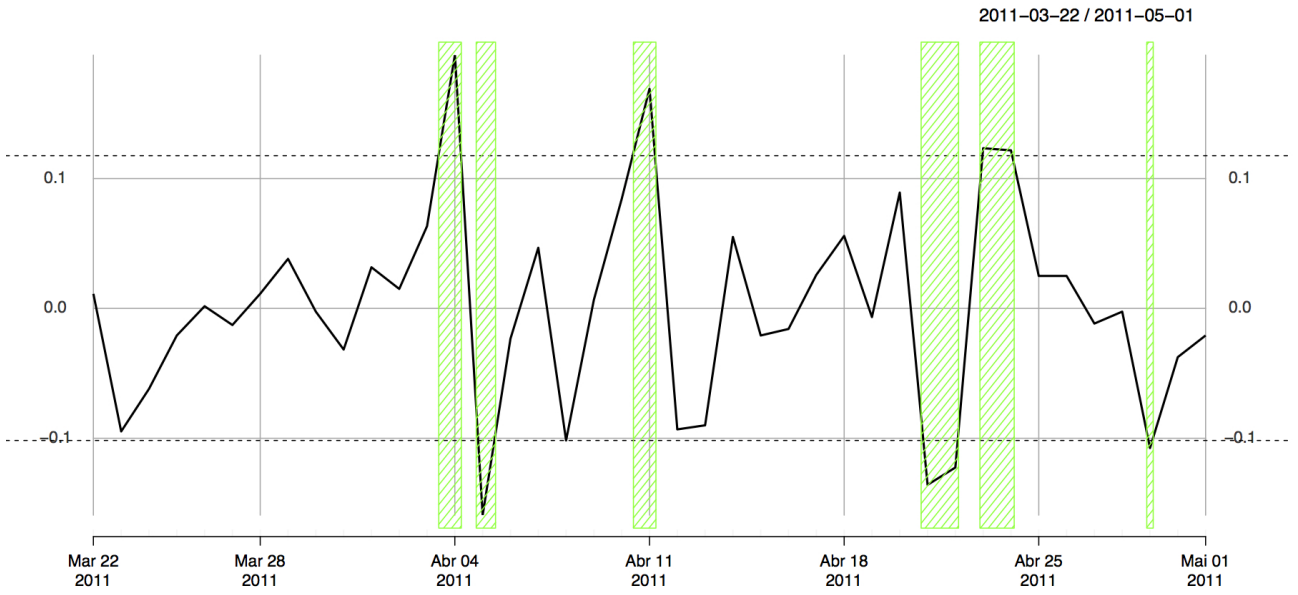
- Time series often exhibit systematic changes in the observed values distribution
- Concept Drift

Temporal Bias: favour cases within the vicinity of apparent regime changes.

Relevance Bins

• Successive observations of the time series where the observed values is either relevant of irrelevant, for the user

•Bins are created using time stamp information and the relevance of the values from the original time series



Proposal

- Improve the predictive accuracy of rare cases in time series forecasting
- Extend resampling strategies for time series forecasting
- Include temporal bias in resampling strategies
- Include temporal and relevance bias in resampling strategies

Assumptions

Assumption #1: The use of resampling strategies significantly improves the predictive accuracy of imbalanced time series forecasting models in comparison to the standard use of out-of-the-box regression tools.

Assumption #2: The use of a temporal bias in resampling strategies significantly improves the predictive accuracy of imbalanced time series forecasting models in comparison to the baseline versions of each respective strategy.

Assumption #3: The use of resampling strategies significantly improves the predictive accuracy of imbalanced time series forecasting models in comparison to the use of ARIMA models.

Resampling with Temporal Bias

- Cases within each relevance bin are not equally relevant
- The most recent cases may portray important information
- Favour these cases

Undersampling w/ Temporal Bias: for each relevance bin where undersampling is applied, the older the example is, the lower is the preference of it being selected for the new training set.

SMOTE w/ Temporal Bias: building on the former, when performing oversampling, the generation of a new synthetic case is done using the most recent k-nearest neighbour.

Resampling with Temporal and Relevance Bias

- A recent case is not necessarily a more relevant one, distribution-wise
- To account for relevance and recency we combine time and relevance as bias

Undersampling w/Temporal and Relevance Bias: the most recent examples with higher relevance are preferred as to staying in the changed data set.

SMOTE w/Temporal and Relevance Bias: building on the former, when performing oversampling, the generation of a new synthetic case is done using the k-nearest neighbour with the highest product of relevance by time position.

Experimental Evaluation

- 24 Datasets (6 Sources)
- 4 Regression Algorithms
- Monte Carlo Estimates (50% for training, 25% for testing)
- Assumption validity: paired comparison using Wilcoxon signed rank tests ($p<0.05$)

Data

ID	Time Series	Data Source	Granularity	Characteristics	% Rare
DS1	Temperature	Bike Sharing [16]	Daily	From 01/01/2011 to 31/12/2012 (731 values)	9.9%
DS2	Humidity				9.3%
DS3	Windspeed				7.8%
DS4	Count of Bike Rentals				13.3%
DS5	Temperature		Hourly	From 01/01/2011 to 31/12/2012 (7379 values)	3.5%
DS6	Humidity				4.8%
DS7	Windspeed				12.5%
DS8	Count of Bike Rentals				17.6%
DS9	Flow of Vatnsdalsa River	Icelandic River [19]	Daily	From 01/01/1972 to 31/12/1974 (1095 values)	21.1%
DS10	Minimum Temperature	Porto weather[1]	Daily	From 01/01/2010 to 28/12/2013 (1457 values)	4.8%
DS11	Maximum Temperature				13.3%
DS12	Maximum Steady Wind				11%
DS13	Maximum Wind Gust				11.1%
DS14	SP	Istanbul Stock Exchange [20]	Daily	From 05/01/2009 to 22/02/2011 (536 values)	16.3%
DS15	DAX				11.4%
DS16	FTSE				9.7%
DS17	NIKKEI				11.6%
DS18	BOVESPA				10.1%
DS19	EU				8.2%
DS20	Emerging Markets				6.8%
DS21	Total Demand	Australian electricity load [21]	Half-Hourly	From 01/01/1999 to 01/09/2012 (239602 values)	1.8%
DS22	Recommended Retail Price				10.2%
DS23	Pedrouços	Water Consumption of Oporto[2]	Half-Hourly	From 06/02/2013 to 11/01/2016 (51208 values)	0.08%
DS24	Rotunda AEP				3.4%

[1] Source: Freemeteo http://freemeteo.com.pt/
[2] Source: Águas do Douro e Paiva http://addp.pt/

Regression Algorithms

ID	Method	R package
LM	Multiple linear regression	stats
SVM	Support vector machines	e1071
MARS	Multivariate adaptive regression splines	earth
RF	Random Forests	randomForest

Evaluation Metrics

- Standard evaluation metrics are not suitable to evaluate rare case prediction tasks (e.g. mse, mae, …)
- We resort to the utility-based framework proposed by Ribeiro [Ribeiro, 2011]
- This framework proposes several evaluation metrics designed for rare case prediction tasks
- We use F-Score, based on the definition of precision and recall for regression proposed by Ribeiro [Ribeiro, 2011]

Results

Model	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	DS10	DS11	DS12
lm	0.027	0.027	0.154	0.219	0.000	0.000	0.027	0.244	0.100	0.161	0.033	0.147
U_B	0.208	0.434	0.307	0.383	0.161	0.069	0.204	0.352	0.099	0.418	0.141	0.382
U_T	**0.210**	0.424	0.308	0.380	0.163	0.070	0.205	0.343	**0.116**	0.416	0.152	0.391
U_TPhi	0.207	**0.439**	**0.318**	**0.389**	**0.170**	**0.082**	**0.225**	**0.434**	0.113	**0.467**	**0.159**	**0.402**
SM_B	0.231	**0.443**	0.330	0.412	0.203	0.146	0.276	0.389	0.118	0.416	**0.191**	0.437
SM_T	0.241	0.416	**0.340**	0.419	0.200	0.170	0.306	0.388	0.166	0.417	0.183	0.439
SM_TPhi	**0.266**	0.423	0.336	**0.421**	**0.206**	**0.177**	**0.319**	**0.436**	**0.167**	**0.461**	0.186	**0.452**
svm	0.107	0.000	0.063	0.082	0.021	0.000	0.027	0.496	0.083	0.196	0.053	0.051
U_B	0.162	**0.256**	0.179	0.230	**0.221**	0.181	0.238	**0.526**	**0.278**	0.394	0.168	**0.261**
U_T	0.175	0.254	0.170	0.246	0.218	0.187	0.244	0.524	0.277	0.393	0.165	0.252
U_TPhi	**0.179**	0.243	**0.198**	**0.260**	0.217	**0.234**	**0.294**	0.512	0.268	**0.464**	**0.186**	0.260
SM_B	0.171	0.283	**0.221**	0.274	**0.235**	0.269	0.313	**0.539**	**0.292**	0.279	0.225	0.315
SM_T	0.214	0.294	0.211	**0.294**	0.225	0.259	0.324	0.535	0.226	0.290	0.230	0.305
SM_TPhi	**0.229**	**0.297**	0.211	0.292	0.223	**0.279**	**0.349**	0.527	0.216	**0.375**	**0.253**	**0.318**
mars	0.044	0.089	0.192	0.213	0.005	0.000	0.044	0.406	0.116	0.172	0.067	0.162
U_B	0.204	**0.299**	0.236	0.341	0.191	0.097	0.228	0.457	0.142	0.393	0.111	0.349
U_T	0.228	0.294	0.239	**0.367**	0.193	0.096	0.232	0.458	0.138	0.396	0.120	0.341
U_TPhi	**0.243**	0.291	**0.282**	0.355	**0.200**	**0.115**	**0.250**	**0.461**	**0.150**	**0.466**	**0.140**	**0.352**
SM_B	0.251	0.369	0.325	0.400	**0.236**	0.184	0.296	0.479	0.144	0.387	0.211	**0.410**
SM_T	0.293	0.361	0.315	0.402	0.219	0.193	0.323	0.494	**0.175**	0.372	0.224	0.397
SM_TPhi	**0.307**	**0.393**	**0.349**	**0.405**	0.223	**0.201**	**0.333**	**0.507**	0.173	**0.430**	**0.229**	0.396
rf	0.010	0.010	0.000	0.198	0.032	0.000	0.060	0.476	**0.150**	0.112	0.043	0.041
U_B	0.142	0.122	0.079	0.260	0.201	0.119	0.222	**0.520**	**0.150**	0.381	**0.102**	**0.164**
U_T	0.133	0.124	0.080	**0.270**	**0.207**	0.118	0.225	0.517	0.143	0.378	0.094	0.159
U_TPhi	**0.148**	**0.131**	**0.088**	0.267	0.203	**0.142**	**0.245**	0.499	0.145	**0.472**	0.096	**0.164**
SM_B	0.151	0.181	**0.096**	0.309	0.129	0.095	0.206	**0.521**	0.156	0.234	0.119	0.203
SM_T	0.169	0.213	0.083	0.323	0.115	0.087	0.220	**0.521**	0.163	0.225	0.133	0.206
SM_TPhi	**0.180**	**0.224**	0.084	**0.329**	**0.138**	**0.125**	**0.257**	0.508	**0.170**	**0.418**	**0.136**	**0.233**
ARIMA	0.015	0.000	**0.158**	0.231	0.000	0.000	0.037	0.184	**0.147**	0.179	0.039	0.137

	DS13	DS14	DS15	DS16	DS17	DS18	DS19	DS20	DS21	DS22	DS23	DS24
lm	0.146	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.218	0.394	0.123	0.417
U_B	0.419	0.059	0.026	0.152	0.057	**0.035**	**0.158**	0.012	**0.219**	0.508	0.168	0.453
U_T	0.426	**0.085**	0.049	**0.187**	0.053	0.030	0.154	**0.032**	**0.219**	**0.509**	0.170	0.455
U_TPhi	**0.445**	0.068	**0.051**	0.161	**0.079**	0.031	0.148	0.024	0.218	0.505	**0.174**	**0.458**
SM_B	0.466	0.145	0.062	**0.204**	**0.183**	0.067	0.197	0.020	**0.219**	**0.519**	0.164	0.456
SM_T	0.462	0.157	**0.119**	0.170	0.135	0.107	0.215	**0.063**	**0.219**	0.351	0.163	0.462
SM_TPhi	**0.476**	**0.164**	0.112	0.183	0.150	**0.117**	**0.218**	0.053	0.218	0.344	**0.168**	**0.472**
svm	0.109	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.216	**0.484**	0.176	0.427
U_B	0.316	**0.109**	0.006	0.099	0.059	0.024	0.061	0.002	**0.218**	0.405	0.270	0.470
U_T	0.302	0.080	0.002	**0.103**	0.060	0.016	0.051	0.004	**0.218**	0.406	0.256	**0.472**
U_TPhi	**0.344**	0.107	**0.031**	0.093	**0.127**	**0.033**	**0.084**	**0.014**	0.217	**0.412**	**0.282**	0.453
SM_B	0.371	**0.205**	0.033	0.144	0.142	0.073	**0.080**	0.007	**0.217**	**0.410**	0.180	0.469
SM_T	0.348	0.161	0.064	0.113	0.160	**0.080**	0.062	0.006	**0.217**	0.305	0.165	**0.477**
SM_TPhi	**0.372**	0.191	**0.074**	**0.153**	**0.187**	0.078	**0.080**	**0.012**	**0.217**	0.324	**0.198**	0.453
mars	0.132	0.018	0.000	0.008	0.000	0.000	0.004	**0.020**	**0.218**	**0.362**	0.155	0.423
U_B	0.372	0.117	0.022	0.080	0.067	0.034	0.026	0.000	**0.218**	0.350	**0.224**	0.474
U_T	0.386	**0.166**	0.029	0.080	0.070	**0.045**	0.014	0.010	**0.218**	0.354	0.221	**0.475**
U_TPhi	**0.391**	0.126	**0.038**	**0.084**	**0.085**	0.034	**0.041**	**0.015**	**0.218**	**0.368**	0.221	0.454
SM_B	0.423	**0.242**	0.098	**0.226**	0.205	0.144	0.136	0.007	**0.218**	**0.345**	0.178	0.473
SM_T	0.414	0.232	0.094	0.196	0.193	0.145	0.179	0.059	**0.218**	0.303	0.164	**0.482**
SM_TPhi	**0.429**	0.240	**0.113**	0.195	**0.217**	**0.179**	**0.199**	**0.063**	0.217	0.322	**0.197**	0.464
rf	0.098	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.215	0.398	0.179	0.429
U_B	0.193	0.036	**0.004**	0.058	0.002	**0.037**	0.023	**0.003**	**0.225**	**0.428**	0.161	0.476
U_T	0.205	**0.048**	0.000	**0.063**	**0.008**	0.028	0.028	0.000	**0.225**	**0.428**	0.170	**0.480**
U_TPhi	**0.208**	0.039	0.003	0.058	0.007	0.033	**0.057**	0.000	0.218	0.409	**0.223**	0.452
SM_B	0.230	**0.084**	0.006	**0.097**	0.060	0.071	**0.050**	0.000	0.217	**0.423**	0.239	0.464
SM_T	0.240	0.075	0.004	0.049	0.061	0.074	0.035	**0.016**	**0.218**	0.388	0.228	**0.479**
SM_TPhi	**0.261**	**0.084**	**0.014**	0.045	**0.089**	**0.100**	0.028	**0.016**	0.217	0.377	**0.242**	0.462
ARIMA	0.146	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.218**	**0.387**	0.148	0.427

Assumption #1

Assumption: “The use of resampling strategies significantly improves the predictive accuracy of imbalanced time series forecasting models in comparison to the standard use of out of the box regression tools.”

	LM	SVM	MARS	RF
U_B	22 (21) / 2 (1)	24 (21) / 0 (0)	23 (19) / 1 (0)	23 (19) / 1 (0)
SM_B	23 (23) / 1 (1)	23 (21) / 1 (0)	23 (21) / 1 (1)	22 (21) / 2 (0)

Win (Sig. Win) / Loss (Sig. Loss)

Assumption #2

Assumption: “The use of a temporal bias in resampling strategies significantly improves the predictive accuracy of imbalanced time series forecasting models in comparison to the baseline versions of each respective strategy.”

	LM.U_B	SVM.U_B	MARS.U_B	RF.U_B
U_T	16 (7) / 8 (1)	11 (1) / 13 (1)	19 (3) / 5 (0)	14 (1) / 10 (2)
U_TPhi	19 (10) / 5 (2)	15 (9) / 9 (6)	19 (8) / 5 (1)	17 (5) / 7 (3)

	LM.SM_B	SVM.SM_B	MARS.SM_B	RF.SM_B
SM_T	14 (7) / 10 (2)	11 (5) / 13 (8)	13 (8) / 11 (4)	15 (8) / 9 (5)
SM_TPhi	18 (12) / 6 (3)	12 (8) / 12 (6)	16 (11) / 8 (2)	17 (13) / 7 (4)

Assumption #3

Assumption: “The use of resampling strategies significantly improves the predictive accuracy of imbalanced time series forecasting models in comparison to the use of ARIMA models.”

Algorithm	Strategy	ARIMA
LM	U_B	22 (21) / 2 (2)
	U_T	23 (23) / 1 (1)
	U_TPhi	22 (22) / 2 (2)
	SM_B	22 (22) / 2 (2)
	SM_T	23 (22) / 1 (1)
	SM_TPhi	23 (22) / 1 (1)
SVM	U_B	23 (18) / 1 (0)
	U_T	24 (18) / 0 (0)
	U_TPhi	23 (21) / 1 (0)
	SM_B	24 (20) / 0 (0)
	SM_T	22 (20) / 2 (0)
	SM_TPhi	22 (22) / 2 (1)
MARS	U_B	21 (19) / 3 (1)
	U_T	21 (19) / 3 (1)
	U_TPhi	21 (20) / 3 (0)
	SM_B	21 (20) / 2 (2)
	SM_T	22 (20) / 2 (2)
	SM_TPhi	22 (21) / 2 (2)
RF	U_B	22 (18) / 2 (2)
	U_T	20 (18) / 4 (1)
	U_TPhi	20 (18) / 4 (1)
	SM_B	20 (19) / 4 (2)
	SM_T	21 (19) / 3 (1)
	SM_TPhi	21 (19) / 3 (2)

Conclusions

- Significant improvement in predictive accuracy of models, focusing on rare and relevant cases of imbalanced time series data;
- The application of resampling strategies is very useful to improve predictive accuracy
- The use of temporal and/or relevance bias further improves the results when compared to the baseline resampling strategies
- The application of resampling strategies provides a significant advantage in comparison to time series focused ARIMA models

Thank you.

Nuno Moniz
nmmoniz@inescporto.pt

Paula Branco
paobranco@gmail.com

Luís Torgo
ltorgo@dcc.fc.up.pt

Code + Presentation @ github:nunompmoniz