Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

SOGOOD 2019:

Prediction of Frequent Out-Of-Hours' Medical Use

Duncan Wallace, Tahar Kechadi

A World Leading SFI Research Centre

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI)under Grant Number SFI/12/RC/2289, co-funded under the European Regional Development Fund

Background

Insight
Centre for Data Analytics

- Out-of-Hours' Care (OOHC) provides medicinal services during periods when family doctors are not available
- Common health platform in many countries
- Large telemedical element

- Cases are treated independent of one another
- Cases are predominantly recorded as free text

Problem Statement

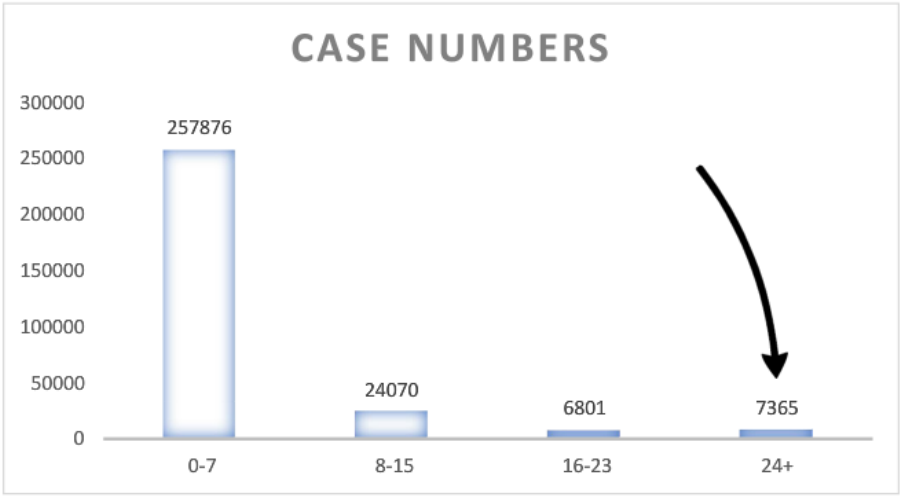
Insight
Centre for Data Analytics

- Prediction of high volume users
- Medical and operational motivations for detecting these patients
- Data predominantly consists of medical domain natural language
- Previous tests had indicated that RNNs featuring LSTM would be suitable for this task
- This research was designed to determine:
  - Whether frequent users could be accurately predicted
  - If a recurrent neural network could outperform other methods
  - If free-text or parameterised data worked better as input

Dataset description

Insight
Centre for Data Analytics

- Target class is outlier
- Heterogeneous dataset: large volume of data, but a lot of sparsity, noise, and missing values.
- Virgin dataset - no previous classification analysis.
- Fairly analogous to data produced by other OOHC bodies (particularly telemedical OOHC).



	mean	std	min	max
dob	1980.64	27.42	1900	2020
month	6.59	3.56	1.00	12
day	15.98	8.83	1.00	31
cases	6.35	23.62	1.00	395
weekday	3.92	2.39	1.00	7
hour	15.21	5.478	1.00	23
minute	28.68	17.25	0.00	59
second	29.41	17.29	0.00	59
case-no	58554.32	23856.08	10001	99999
Cons_Time_Taken	8.08	8.7	0.00	356

Baseline analysis

Insight
Centre for Data Analytics

- Multiple 'traditional' machine learning approaches adopted in relation to both parameterised and free-text data

Parameterised data for classification

	KNN	Naive Bayes	SVM
Accuracy	0.47 ±0.061	0.79 ±0.28	0.50 ±0.05
PPV	0.34 ±0.06	0.26 ±0.33	0.74 ±0.06
NPV	0.42 ±0.06	0.8 ±0.29	0.51 ±0.05
TPR	0.009 ±0.007	0.24 ±0.01	0.02 ±0.0005

Classification using bag-of-word model

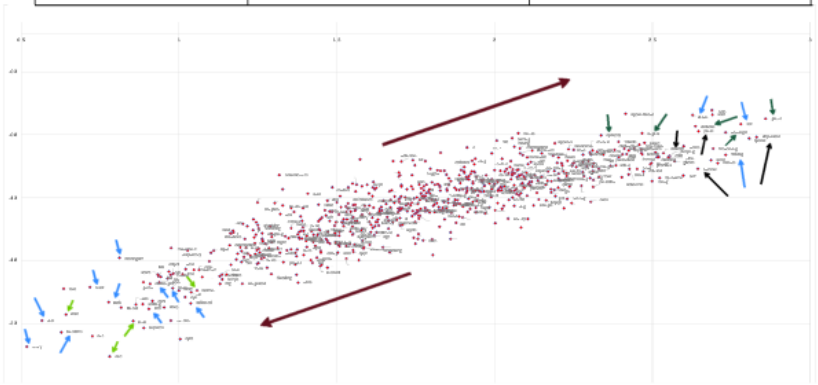
	Naive Bayes	SVM	RF
Accuracy	0.26 ±0.51	0.66 ±0.385	0.68 ±0.045
PPV	0.89 ±0.53	0.52 ±0.459	0.37 ±0.015
NPV	0.24 ±0.019	0.49 ±0.412	0.69 ±0.044
TPR	0.013 ±0.001	0.01 ±0.0049	0.006 ±0.001

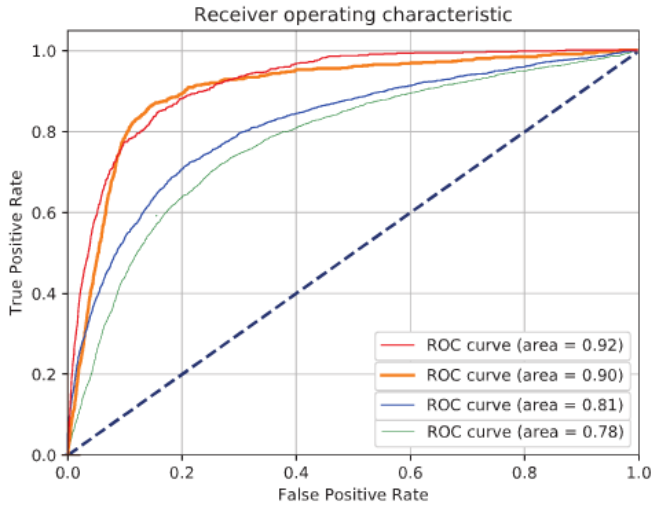
Classification

Insight
Centre for Data Analytics

- Classification compared variable input length and channel number. Also input with patient details (age, sex, etc.) added as features versus noise input.

			LSTM			
			Features		Noise	
T	Length	C	PPV	NPV	PPV	NPV
24	100	100	0.71±0.006	0.76±0.005	0.75±0.033	0.75±0.047
24	100	200	0.73±0.021	0.75±0.003	0.72±0.047	0.76±0.013
24	200	100	0.71±0.018	0.72±0.012	0.54±0.21	0.89±0.081
24	200	200	0.78±0.098	0.63±0.21	0.38±0.33	0.91±0.042
50	100	100	0.88±0.009	0.78±0.008	0.86±0.012	0.88±0.014
50	100	200	0.85±0.017	0.83±0.026	0.88±0.019	0.84±0.025
50	200	100	0.80±0.083	0.69±0.23	0.79±0.015	0.71±0.17
50	200	200	0.89±0.048	0.75±0.165	0.82±0.14	0.72±0.21





Conclusion

Insight
Centre for Data Analytics

- Frequent users are a poorly understood sub-set of patients which provide significant challenges for classification
- These challenges include their demographic variance, and ill-defined medical definition.
- Outlier detection of this nature tends to suffer from high variance. Matters are nohelped by the recording techniques common to many OOHC providers.
- RNN with gating did well in accurately predicting these patients, with results appearing to differentiate between chronic and acute patient episodes.