Deep learning applied to mobile phone data for Individual income classification

Pål Sundsøy[1*], Johannes Bjelland[1], Bjørn-Atle Reme[1], Asif M.Iqbal[1] and Eaman Jahani[2]
[1]Telenor Group Research, Snarøyveien 30,1331 Fornebu, Norway
[2]Massachusetts Institute of Technology, The Media Laboratory, USA
[*]pal-roe.sundsoy@telenor.com

Abstract— **Deep learning has in recent years brought breakthroughs in several domains, most notably voice and image recognition. In this work we extend deep learning into a new application domain - namely classification on mobile phone datasets. Classic machine learning methods have produced good results in telecom prediction tasks, but are underutilized due to resource-intensive and domain-specific feature engineering. Moreover, traditional machine learning algorithms require separate feature engineering in different countries. In this work, we show how socio-economic status in large de-identified mobile phone datasets can be accurately classified using deep learning, thus avoiding the cumbersome and manual feature engineering process. We implement a simple deep learning architecture and compare it with traditional data mining models as our benchmarks. On average our model achieves 77% AUC on test data using location traces as the sole input. In contrast, the benchmarked state-of-the-art data mining models include various feature categories such as basic phone usage, top-up pattern, handset type, social network structure and individual mobility. The traditional machine learning models achieve 72% AUC in the best-case scenario. We believe these results are encouraging since average regional household income is an important input to a wide range of economic policies. In underdeveloped countries reliable statistics of income is often lacking, not frequently updated, and is rarely fine-grained to sub-regions of the country. Making income prediction simpler and more efficient can be of great help to policy makers and charity organizations – which will ultimately benefit the poor.**

Deep learning; Mobile phone data; poverty;household income ; Big Data Analytics; Machine learning; Asia, Mobile network operator; metadata; algorithms

I. INTRODUCTION

Recent advances in Deep Learning [1][2] have made it possible to extract high-level features from raw sensory data, leading to breakthroughs in computer vision [9][10][11] and speech recognition [12][13]. It seems natural to ask whether similar techniques could also be beneficial for useful prediction tasks on mobile phone data, where classic machine learning algorithms are often under-utilized due to time-consuming country and domain- specific feature engineering [6].

Our work investigates how we can separate individuals with high and low socio-economic status using mobile phone call detail records (CDR). Finding good proxies for income in mobile phone data could lead to better poverty prediction – which could ultimately lead to more efficient policies for addressing extreme poverty in hardest hit regions.

With this in mind, we perform a large-scale country-representative survey in a low HDI Asian country, where household income is collected from approximately 80 000 individuals. The individual records are de-identified and coupled with raw mobile phone data that span over 3 months. This dataset allow us to build a deep learning model, as well as a benchmarking model using custom feature engineering and traditional data mining algorithms.

From the household income we derive two binary classifiers (1) below or above median household income and (2) below or above upper poverty level. The income threshold for poverty level is derived from official statistics and based on average national household income and household size. Our survey classifies participants into 13 household income bins – where bin 1 and 2 correspond to below upper poverty level.

The rest of this paper is organized as follows: In section 2 we describe the features and models used for benchmarking our deep learning approach. Section 3 describes the deep learning approach itself. In section 4 we compare the results of the two approaches. Finally, we draw our conclusions in section 5.

II. BEST PRACTICE – TRADITIONAL DATA MINING

This section describes the features and the standard machine learning algorithms used as our benchmark. The data preparation phase in the data mining process is a tedious task, that requires specific knowledge about both the local market and the various data channels as potential sources for input features. Typically the data warehouse architecture and the data format vary between the operators and third-party software packages, making it hard to create generalizable feature-sets.

A. Features

We build a structured dataset consisting of 150 features from 7 different feature families, see Table 1. The features are

custom-made and range from basic phone usage, top-up pattern, social network and mobility to handset usage. The features include various parameters of the corresponding distributions such as weekly or monthly median, mean and variance.

TABLE 1 SAMPLE FEATURES USED IN TRADITIONAL MACHINE LEARNING BENCHMARKS.

Feature family	Feature examples
Basic phone usage	Outgoing/incoming voice duration, sms count etc.
Top-up transactions	Spending speed, recharge amount per transaction, fraction of lowest/highest recharge amount, coefficient of variation recharge amount etc
Location/mobility	Home district/tower, radius of gyration, entropy of places, number of places etc.
Social Network	Interaction per contact, degree, entropy of contacts etc.
Handset type	Brand, manufacturer, camera enabled, smart/feature/basic phone etc
Revenue	Charge of outgoing/incoming SMS, MMS, voice, video, value added sevices, roaming, internet etc.
Advanced phone usage	Internet volume/count, MMS count, video count/duration, value added services duration/count etc.

A. *Models*

Ensemble methods have been proven as powerful algorithms when applied to large-scale telecom datasets [14][6]. They combine predictions of several base classifiers built with a given learning algorithm in order to improve robustness over a single classifier. We investigate two different state-of-the-art ensemble methods:

1) Random forest, where we build several independent classifiers and average their predictions, thus reducing the variance. We choose grid search to optimize the tree size.

2) Gradient boosting machines (GBM) where the base classifiers are built sequentially. The algorithm combines the new classifier with ones from previous iterations in an attempt to reduce the overall error rate. The main motivation is to combine several weak models to produce a powerful ensemble.

In our set-up, each model is trained and tested using a 75/25 split. For the response variable related to poverty level we introduce misclassfication cost. Since few people are below the poverty level (minority class), a naive model will predict everyone above poverty line. We therefore apply a misclassification cost to the minority class to achieve fewer false positives and adjust for the ratio between the classes.

III. DEEP LEARNING

In this section we describe the input data and the structure of our deep learning model.

B. *Features*

Classification results of the traditional learning algorithms are inherently limited in performance by the quality of the extracted features [7]. Deep learning can instead reproduce complicated functions that represent higher level extractions, and replace manual domain-specific feature engineering.

Earlier studies have shown a good correlation between location/human mobility and socio-economic levels [15-17]. Using this as a motivation we build a simple vector whose length corresponds to the number of mobile towers (8100 dimensions), and the vector elements correspond to the mobile phone activity at the given tower – shown in Table 2.

TABLE 2 INPUT VECTORS TO DL ALGORITHM BEFORE NORMALIZATION

Hashed Phone number	Tower 1	Tower2	….	Tower8100	Below/above poverty level
1	0	67	….	16	0
2	7	0	….	0	1
.	.	.	….	.	.
.	.	.		.	.
80K	0	9	….	6	0

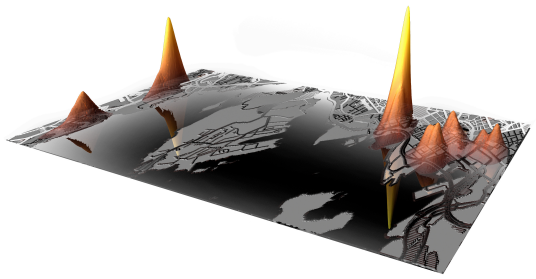

FIGURE 1 LOCATION ACTIVITY MAP. THE MAP PROVIDES INFORMATION ABOUT EACH USERS' TOWER ACTIVITY.

A visual representation of Table 2 can be seen in Fig 1, where each subscriber is represented by a location density map.

A. *Models*

We use a standard multi-layer feedforward architecture where the weighted combination of the n input signals is aggregated, and an output signal $f(\alpha)$ is transmitted by the connected neuron. The function f used for the nonlinear

activation is rectifier $f(\alpha) \approx \log(1+e^{\alpha})$. To minimize the loss function we apply a standard stochastic gradient descent with the gradient computed via back-propagation. We use dropout [4][5] as a regularization technique to prevent over-fitting. For the input layer we use the value of 0.1 and 0.2 for the hidden layers. Dropout secures that each training example is used in a different model which all share the same global parameters. This allows the models to be averaged as ensembles, improving generalization and preventing overfitting. The split between train and test set, as well as the introduced misclassification cost for poverty level, are similar to the benchmark models.

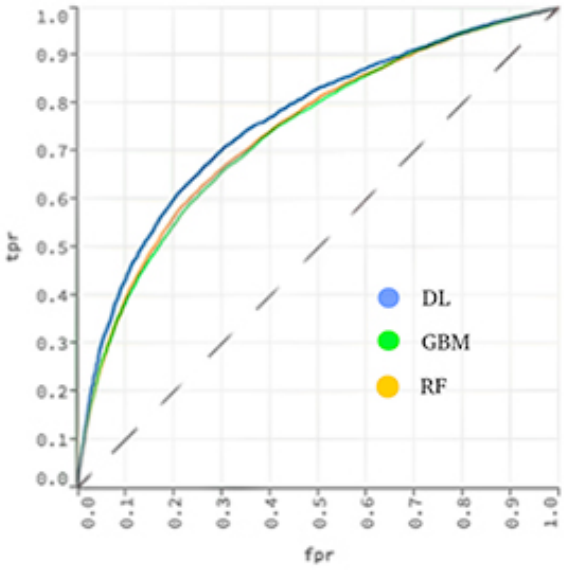

FIGURE 2 AUC ON TEST SET SHOWING THE TRUE POSITIVE VS FALSE POSITIVE RATE FOR DEEP LEARNING (DL), GRADIENT BOOSTING MACHINES (GBM) AND RANDOM FOREST (RF).

TABLE 3 MODEL PERFORMANCE (AUC) ON TEST SET

Model	AUC test set Below/above poverty level	AUC test set Below/above median income
DL	**0.74**	**0.77**
GBM	0.68	0.71
RF	0.69	0.72

IV. RESULTS

Our deep neural network is trained to separate individuals with high and low socio-economic status. We evaluate our models' performance with AUC [2][3] achieving 77% and 74% AUC on test set when the classifier is predicting above/below median household income and above/below poverty level respectively (Fig 2, Table 3). The corresponding AUC on the train sets are respectively 80% and 77%, showing that our model does not overfit significantly. Our results indicate that the DL model achieves a higher performance than those of multi-source RF and GBM models, which vary between 71-72% and 68-69% for above/below median household income and poverty level respectively.

Next, we feed the RF and GBM models with the same representation as fed into the DL algorithm - achieving RF performance of AUC 64% and GBM performance of AUC 61% . The performance of these models greatly suffers as the number of input features increase without increasing the training size. We conclude that given a fixed training sample, the traditional models are not able to learn a complicated function that represents higher level extractions, but perform better when using manual domain-specific feature engineering.

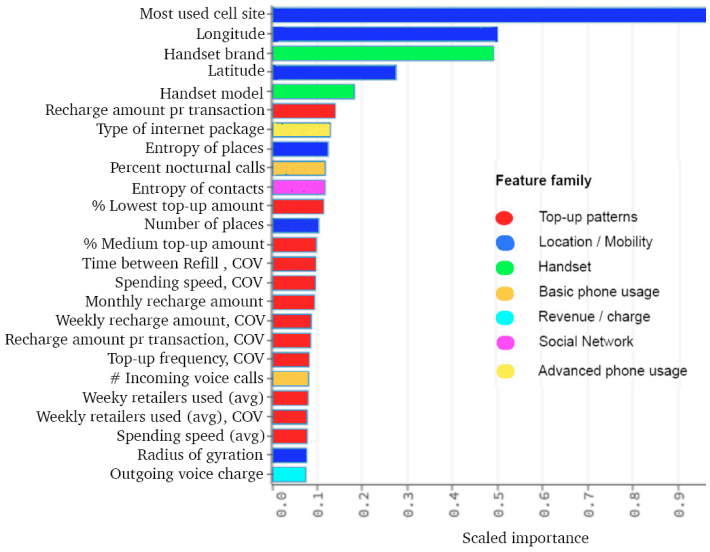

FIGURE 3 TOP FEATURES IN RF MODEL COLORED BY THEIR RESPECTIVE FEATURE FAMILY.

A posteriori inspection of the top features selected by our traditional models leads to some interesting qualitative insights. We observe a similar pattern in the top features selected by the RF and GBM model. Figure 3 shows the top features in our RF model. We notice the importance of three feature families:

1) Location dynamics: Where the user spends most of his time is a good signal of his income. This indicates that our models have detected regions of low economic development status.

2) The handset brand: In the country of our study, minimal and more affordable handset brands are very popular among the lower income quantiles, while expensive smartphones are considered as a huge status symbol.

3) The top-up pattern: Interestingly, the recharge amount per transaction is more predictive than the total recharge amount. We observe that individuals from the lower income quantiles usually top-up with lower amounts when they first fill up their account.

V. CONCLUSION

In order to predict household income based on mobile phone communication and mobility patterns, we implemented a multi-layer feedforward deep learning architecture. Our approach introduces a novel data representation for learning neural networks on real CDR data.

Our approach suggests that multi-layer feedforward models are an effective tool for predicting economic indicators based on mobile communication patterns. While capturing the complex dependencies between different dimensions of the data, deep learning algorithms do not overfit the training data as seen by our test performance.

Furthermore, our deep learning model, using only a single dimension of the data in its raw form, achieves a 7% better performance compared to the best traditional data mining approach based on custom engineered features from multiple data dimensions. Even though such an automated approach is time-saving, many of the classic machine learning approaches have the advantage of being interpretable. However, since a large portion of a data mining process is data preparation, there is a big demand to automate this initial step.

As future work, we would like to investigate the performance implications of including temporal aspects of raw CDRs in our models and the data representation. In addition, we will work on finding a general representation of telecom data that can be used for various prediction tasks.

An important application of this work is the prediction of regional and individual poverty levels in low HDI countries. Since our approach only requires de-identified customer and tower IDs, we find this method more privacy preserving compared to traditional data mining approaches where the input features may reveal sensitive information about the customers.

REFERENCES

[1] Yann LeCun, Yoshua Bengio & Geoffrey Hinton. Deep Learning, Nature 521 436-444 (28 May 2015)

[2] P.Baldi et al. Searching for exotic particles in high-energy physics with deep learning, Nature Communications 5, article 4308

[3] AUC:A better measure than accuracy in comparing learning algorithms, Ling C, Advances in Artificial Intelligence LNCS Vol 2671, 2003, pp 329-341

[4] G. Hinton et. al. (2012) Improving Neural Networks by Preventing Co-adaptation of Feature Detectors. arXiv preprint arXiv:1207.0580

[5] G. Dahl et. al. (2013) Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout. ICASSP, 8609-8613.

[6] P.Sundsøy, J.Bjelland, A.Iqbal, A.Pentland, Y.A.Montjoye, Big Data-Driven Marketing: How machine learning outperforms marketers' gut-feeling, Social Computing, Behavioral-Cultural Modeling & Prediction. Lecture Notes in Computer Science Volume 8393, 2014, pp 367-374

[7] Arel I,Rose DC,Karnowski TP, Deep machine learning – A new frontier in artificial intelligence research. IEEE computational intelligence magazine 5,13-18 (2010).

[8] Bengio Y, Learning deep architectures for AI. Foundations and trends in machine learning 2 (2009).

[9] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pages 1106–1114, 2012

[10] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In Proc. International Conference on Computer Vision and Pattern Recognition (CVPR 2013). IEEE, 2013.

[11] Volodymyr Mnih. Machine Learning for Aerial Image Labeling. PhD thesis, University of Toronto, 2013.

[12] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. Audio, Speech, and Language Processing, IEEE Transactions on, 20(1):30 –42, January 2012.

[13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. In Proc. ICASSP, 2013

[14] Namhyoung K,Kyu J, Yong.K, A New Ensemble Model for Efficient Churn Prediction in Mobile Telecommunication, 2012 45th Hawaii international conf on system sciences

[15] V.Frias-Martinez,J.Virseda,E.Frias-Martinez. Socio-economic levels and human mobility, Qual Meets Quant Workshop – QMQ 2010 at the int.conf on information & communication technologies and development, ICTD (2010).

[16] Rubio, A., Frias-Martinez, V., Frias-Martinez E. & Oliver, N. (2010, March). Human Mobility in Advanced and Developing Economies: A Comparative Analysis, in AAAI Spring Symposia Artificial Intelligence for Development, AID, Stanford, USA.

[17] Eagle, N., Macy, M. & Claxton, R. (2010). Network Diversity and Economic Development, Science 328 2029-1030.