Perspectives on Computational Research: Term Paper Re-write

Laurence Warner

lpwarner@uchicago.edu

Draft of Paper

Who on Earth doesn't own a Smartphone in 2018? Predicting technology uptake with sociological variables.

Abstract

There is a trade-off between using explanatory models which give a clear picture, but don't provide much help in predicting an outcome, with machine learning models, which are black-box models which are very powerful at prediction. Looking at the sociological question of smartphone ownership, I firstly suggest that weakness of the usage gap hypothesis in explaining technology uptake. Secondly, I use more sophisticated machine learning models to show that smartphone ownership can be predicted with high degrees of accuracy.

Keywords: Technology, Demographics, Machine Learning, Prediction

Introduction

Cell phones are a technological device that has rapidly changed society in a unique way. It is projected that by 2022, 90% of all mobile subscriptions will be for Internet-enabled smartphones which are already in the majority (Ericsson, 2016). In the United States, more time is spent on digital activity on smartphones than on computers (ComScore, 2015).

However, technological uptake does not happen in a socioeconomic bubble. The uptake of new technology is biased depending on demographic factors.

Like many technologies before it, internet-disabled cell phones ('dumb phones'), are gradually dying out of usage. However, like some technologies such as the vinyl record, older technology can sometimes make a comeback based on fashion. For example, consider the growing trend of 'digital detox'(https://www.themuse.com/advice/the-detox-you-need-to-go-on-now, soundcloud.com/flipphonediaries)

This may muddy the water of clear linear demographic trends. Hence, it will be interesting to compare these methods to more non-linear machine learning models.

We approach the problem from a predictive standpoint. Given an individual's demographic characteristics, how accurately can we predict the probability of them owning a smartphone.

Literature Review

The primary theoretical and empirical inspiration is Tsetsi et al. (2017): "Smartphone Internet access and use: Extending the digital divide and usage gap". They introduce the concept of the usage gap. This is an adaptation of the sociological theory of the knowledge gap, which posits that that knowledge, like other forms of wealth, is differentially distributed throughout a social system. So it states that technology usage is also distributed differentially.

They also use the Pew Internet survey - but from 2012. However, they did not use any machine learning techniques, so this will be a novel contribution.

Marler (2018): "Mobile phones and inequality: Findings, trends, and future directions" finds that socioeconomic effects are prevalent. But interestingly, low socioeconomic groups are often prone to having smartphones, but no other internet-devices - leading them to be "smartphone dependent". Is the income effect lower than we might expect then?

Andone et al. (2016) "How age and gender affect smartphone usage" use an alternative dataset: their own app "Menthal" to analyze smartphone usage. This data is not available - but I feel that survey data is superior because it represents the entire nation. Non-smartphone users (i.e. 20% of the nation) are not captured by their smartphone app.

Why might it be important to understand who uses a smartphone? As well as the economic funcitons already discussed, there is a growing mental health aspect. Samaha et al. (2016) suggests, smartphone usage is increasingly being linked with psychological dependance issues.

Data

Source

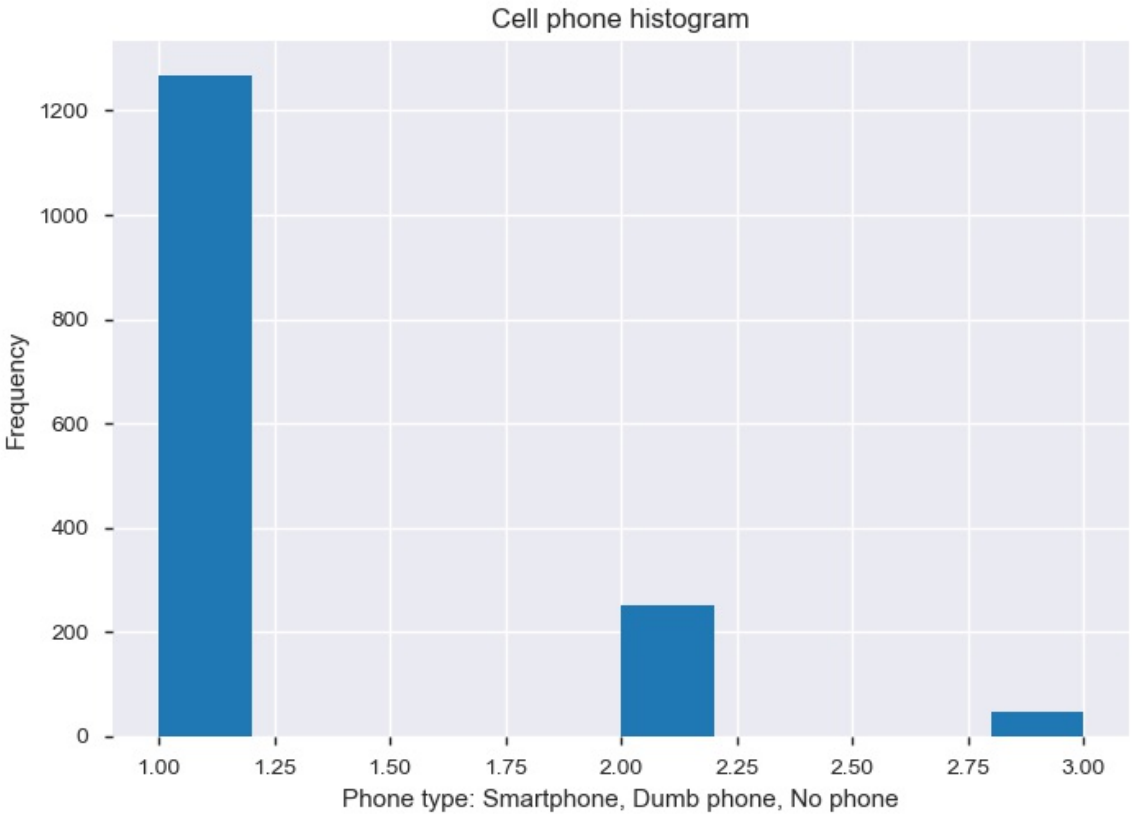
Pew Research Center's Internet Project Core Trends Survey, 2018. It can be accessed freely online here: http://www.pewinternet.org/datasets/

Interviews with a nationally representative sample of 2,002 adults were conducted between January 3-10 2018. The target population for the study is non-institutionalized persons age 18 and over, living in the US. Most of the interviews were conducted using cellphones (n = 1,502) with the remainder conducted using landlines (n=500); both groups were included in the final sample. According to the Pew Research Center, the landline sample was collected using a proportional sample based on listed telephone households. The cellphone sample was selected systematically from dedicated wireless numbers. This dataset contains questions about social media use in 2018 and attitudes toward the internet and whether Americans think it's good or bad for society. Random digit dialing was used to collect survey responses and the final sample was weighted to represent the American adult population. The sample response rate was 11%.

Cleaning & Descriptive Statistics

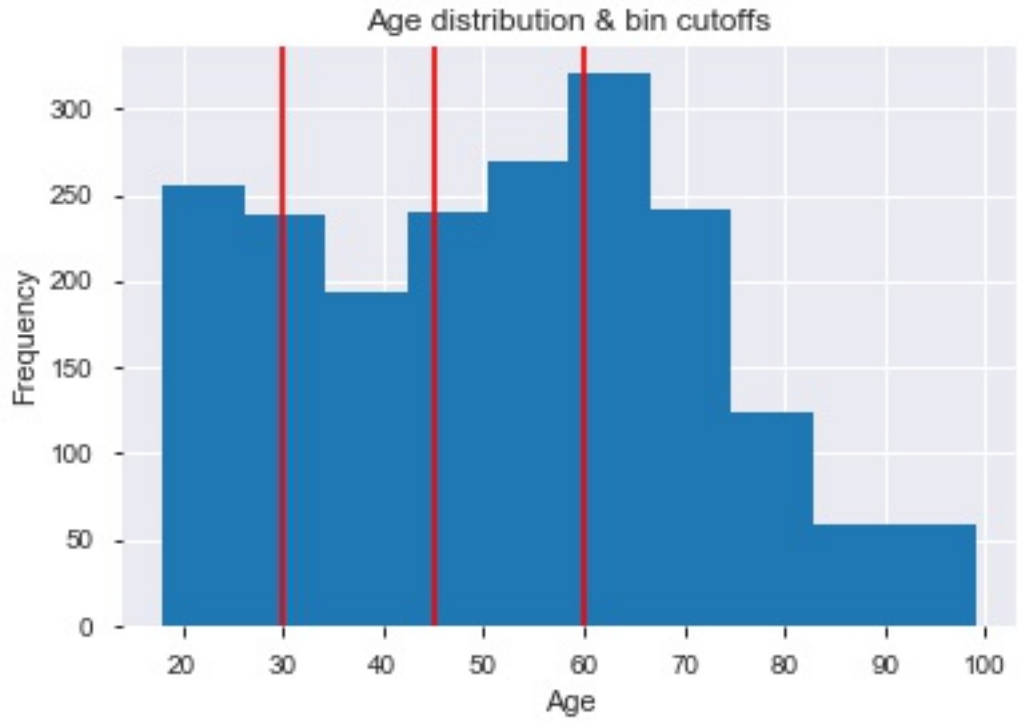
The dataset was cleaned using Python code, including the Pandas library. I chose to discard rows which contained answers including 'Don't Know' or 'Refused to answer'. This reduced the size of the dataset to n = 1567, which is easily enough to run machine learning models reliably (usually a minimum of around 500 is required).

Regarding the target variable, I firstly constructed a ternary variable by inferring that respondents who were not asked the question "Is your cell phone a smartphone, or not?" were from the landline sample and responded 'no' to the previous question: "do you have a cell phone, or not?" The distribution of my constructed variable is displayed below:



To predict using binary classifier methods, I combined folks with a dumb phone and no phone into a no smartphone category.

Regarding the predictor variables, I ensured that each variable was categorical. For example, I categorized age using the following cut-offs:



Methods & Results

Traditional Explanatory Analysis

Before exploring predictive machine learning models, to understand the dataset better, I firstly explore an explanatory analysis using traditional binary dependent variable models. I choose logistic regression as the most widely used and easily interpretable.

Table 1 below shows the regression output:

	coef	std err	z	P>|z|
sex	-2.7108	0.660	-4.105	0.000
age	0.0741	0.148	0.500	0.617
educ	0.9605	0.083	11.572	0.000
hisp	-0.2292	0.049	-4.707	0.000
inc	0.4006	0.272	1.470	0.141
race	-0.2983	0.035	-8.575	0.000

Contrary to the usage gap theory, it is suggested that age and income are not statistically significant variables (p-values absolutely greater than 0.05).

The benefit of this approach is the interpretability of the coefficients. For example, the statistically significant negative coefficient on sex (male = 1, female = 2) suggests that being female reduces the likelihood of eschewing the smartphone (smartphone = 0, no smartphone = 1).

However, this model does not help greatly if tasked with predicting a new individual's tech uptake.

Machine-learning based prediction

The linear model used suggests that the demographic influences on smartphone usage are not simple. For the prediction problem, we will need to build a model which can predict whether an individual will have a smartphone.

I use a 70/30 train-test split: meaning the model is only built on 70% of the data.

In terms of model evaluation, we will focus on accuracy as we are interested in how often we can correctly identify an individual's choice of phone.

Firstly, we use a logistic regression classifier in this new context. I give an example of a classification report:

	precision	recall	f1-score	support
1	0.87	0.97	0.92	390
2	0.67	0.32	0.43	81
avg / total	0.84	0.86	0.83	471

This can be interpreted as follows: from the 390 test case which actually had smartphones, they were correctly predicted 87% of the time.

Also, we can infer that 390/471 = 83% of test set own smartphones. Hence, baseline accuracy is 83% as this would be the accuracy in the trivial case of predicting everyone to have a smartphone.

This logistic regression model underperforms that baseline at 82%.

I show a table here of the accuracy scores for different models:

	accuracy	model
0	81.66	LR
1	90.51	DT
2	90.51	RF
3	81.84	SVM
4	81.84	KNN

As can be seen Decision Tree is the most accurate in this case. Random Forest is the same, because in this instance they are identical models.

The fact that decision tree dominates linear models like Logistic Regression and localised models like K Nearest Neighbors is suggestive that the data is highly non-linear. In other words, there are many anomalies (for example rich, young people eschewing smartphones).

Conclusion and Future Work

In conclusion, this recent data casts some doubt on the Tsetsi's usage gap hypothesis, suggesting that the link between income and tech uptake is not as clear as they thought, in line with Marler's finding.

As hoped, machine learning techniques can outperform the baseline for the prediction problem. In particular, a decision tree is the most accurate of the models considered. This supports my surmise that the data is more complex than considered by Tsetsi etc.

In terms of future direction, it would be interesting to engage in a hybrid of the two approaches discussed: accurate predictive models, but with some human interpretability. The most promising avenue is LIME (Locally Interpretable Model-Agnostic Explanations). However, the current state of the software is not ideal for coping with a dataset like this: with ordered numerical data.

References

Tsetsi, Eric, and Stephen A. Rains. "Smartphone Internet access and use: Extending the digital divide and usage gap." Mobile Media & Communication 5, no. 3 (2017): 239-255.

Andone, Ionut, Konrad Błaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. "How age and gender affect smartphone usage." In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pp. 9-12. ACM, 2016.

Marler, Will. "Mobile phones and inequality: Findings, trends, and future directions." New Media & Society (2018): 1461444818765154.

Samaha, Maya, and Nazir S. Hawi. "Relationships among smartphone addiction, stress, academic performance, and satisfaction with life." Computers in Human Behavior 57 (2016): 321-325.

Afterword

Computer programming can be explored in the draft.ipynb notebook in the same repository.

This document was crafted using Markdown. Markdown is a lightweight markup language with plain text formatting syntax. Although it is not the standard in the academic community, it is highly powerful for creating papers with multimedia elements like this, and I thoroughly

recommend its usage.