Cellular Market Research

Tam N. Nguyen

Abstract

This project offers a quick look at the possibility of identifying the potentials of a cellphone market in anywhere in the world by using just a few socio-economic indicators. The final result is a machine learning model giving reasonable predictions with a reasonable error rate.

Motivation

Mobile devices are quickly becoming more affordable and more powerful. They are becoming the key piece in various business models such as car hailing, home delivery, personal payment systems, online retails, personalized education, and so on. Recognizing emerging cellular communication markets is extremely beneficial to active players in various industries. Predicting the potentials of a cellular market is even more important.

Dataset(s)

The original dataset is the World Development Indicators Dataset:

(https://data.worldbank.org/data-catalog/world-development-indicators)

- 5656458 data entries, 6 columns
- 247 countries, over 1300 indicators
- From 1995 to 2015
- Published by the World Bank

Data Preparation and Cleaning

Trim up data

- Limit time frame from 1995 to 2014 (almost no cell phone subscription prior to 1995). No problem with data cleaning since the data set has been cleaned (for publishing)

Correlate data

- Picked around 50 indicators from 13xx indicators
- Problem: Highest correlation scores may not mean the best indicators
- Problem: unequal spread of collected data across indicators and countries

Prepare datasets for ML model

- Split data set to train and test data sets

Research Question(s)

1. What are the potential indicators for a good, fast growing cellular communication market?
2. What are the top 10 cellular communication markets as of 2015?
3. Is it possible to build a model to predict the growth of any cellular communication market in the world by using just a few common economic/social indicators?

Methods

- Identified “Mobile cellular subscriptions” as the key indicator
- Automatically sweep through World Bank’s “World Development Indicators” database to find indicators that correlate well with the key indicator (>0.85/1)
- Pick top 10 indicators based on correlation scores, number of correlated data entries, and diversity
- Identify top markets based on existing statistics and verify if the chosen indicators do correlate well with “Mobile cellular subscriptions” rates
- Built a decision tree regression model based on the cellular subscription rates and the chosen indicators. Measure the model’s accuracy

* JupyterNotebook Source code:

https://github.com/genterist/DataScience/blob/master/CellphoneMarkets/Cellphone_Market_Research.ipynb

TOP MARKETS BY SUBSCRIPTION GROWTH

1. China - CHN 4.697291e+08
2. India - IND 2.908135e+08
3. United States - USA 1.873658e+08
4. Russian Federation - RUS 1.050839e+08
5. Brazil - BRA 1.044052e+08
6. Indonesia - IDN 9.796699e+07
7. Japan - JPN 9.049884e+07
8. Germany - DEU 6.486960e+07
9. Italy - ITA 6.200658e+07
10. United Kingdom - GBR 5.334796e+07

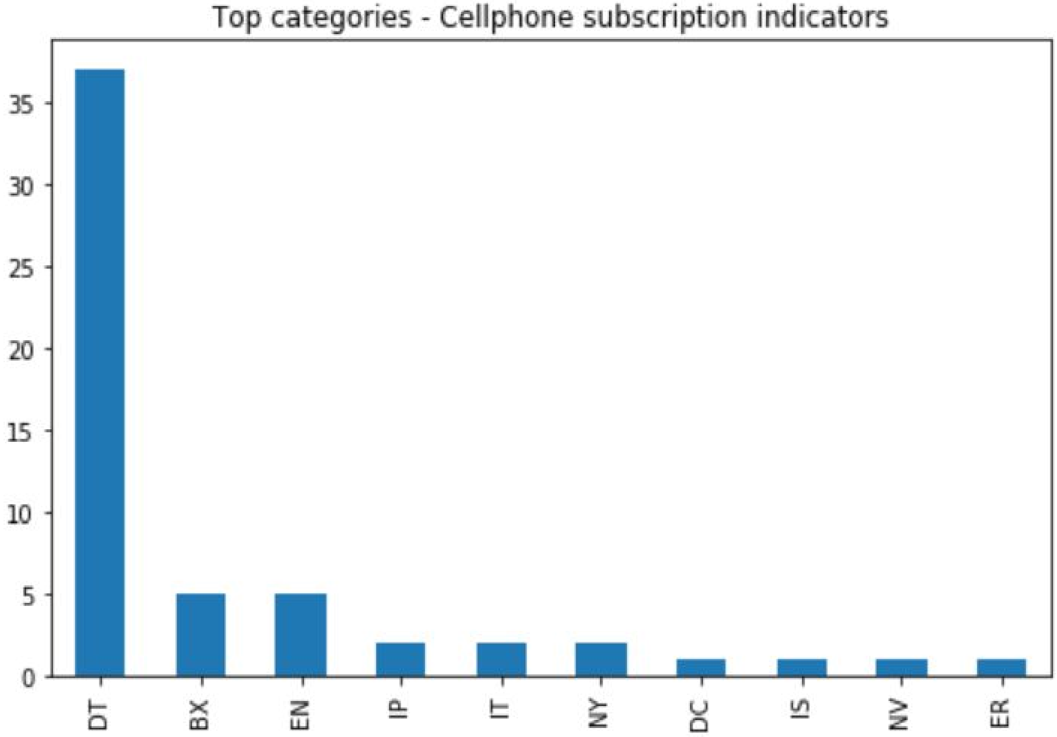
Top Countries by cellphone subscriptions (2014)

1. China - CHN - 1.286093e+09
2. India - IND - 9.440087e+08
3. Indonesia - IDN - 3.190000e+08
4. United States - USA - 3.174438e+08
5. Brazil - BRA - 2.807288e+08
6. Russian Federation - RUS - 2.210304e+08
7. Japan - JPN - 1.526957e+08
8. Nigeria - NGA - 1.389603e+08
9. Vietnam - VNM - 1.361481e+08
10. Pakistan - PAK - 1.357620e+08

Top 10 categories of indicators

x axis : category code
y axis : number of indicators in a category that appear in the list of top 58 indicators

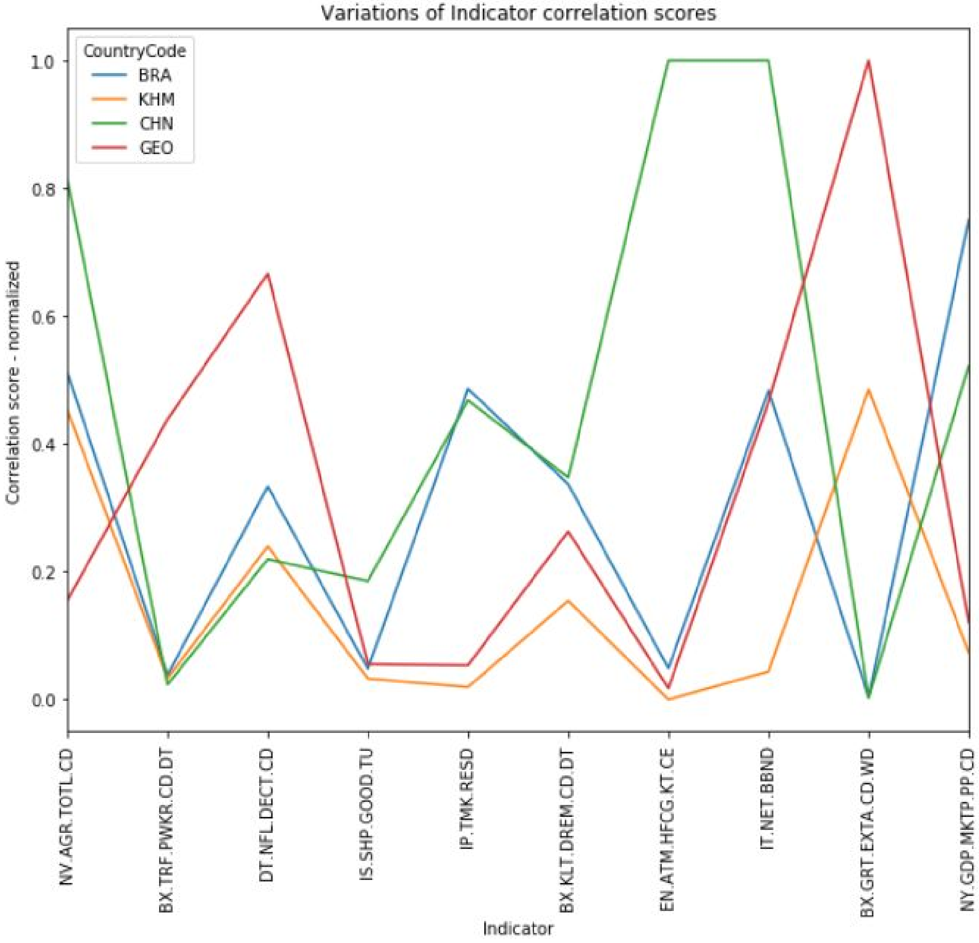
DT : external debt
BX : balance of payments relating to exports
EN : general environment
IP : intellectual property
NY : National accounts relating to income
IT : telecommunication infrastructure
IS : transportation infrastructure
DC : Debts relating to aid flows from DAC
ER : Environment resources
NV : National accounts - value added



Variations of Indicators' correlation scores

Clearly, the overal correlation shape considering all 10 indicators varies greatly from country to country. With the case of Personal remittances, received (BX.TRF.PWKR.CD.DT), while it will not help much in determining the potentials of cellphone market in China, Cambodia and Brazil (correlation score is low), it can be very well be used for Georgia.

So for a machine learning model to be used to predict the potentials of cellphone market in any country in the world, it is better to have a set of diverse features from various indicator categories.



Top Indicators

1. Agriculture, value added : NV.AGR.TOTL.CD
2. Personal remittances, received : BX.TRF.PWKR.CD.DT
3. Net flows on external debt, total : DT.NFL.DECT.CD
4. Container port traffic : IS.SHP.GOOD.TU
5. Trademark applications, direct resident : IP.TMK.RESD
6. Primary income on FDI, payments : BX.KLT.DREM.CD.DT
7. HFC gas emissions (thousand metric tons of CO2) : EN.ATM.HFCG.KT.CE
8. Fixed broadband subscriptions : IT.NET.BBND
9. Grants, excluding technical cooperation : BX.GRT.EXTA.CD.WD
10. GDP, PPP : NY.GDP.MKTP.PP.CD

Machine Learning Model Built

- Based on Decision Tree Regression algorithm
- RMSE = 6.4e+07
- Test set mean: 7.6e+07
- Test set max: 1.1e+09

The error rate of 6.4e+07 while using the Decision Tree regression model is acceptable considering this model only uses 10 indicators in order to predict the potentials of **any** cellphone market of **any** country in the world.

Limitations

Due to the time constraint and limited scope, this project has numerous limitations:

- Indicator selection can be further optmized for a spefic region. A larger number of selected indicators can be made
- More sophisticated Machine Learning model can be used (such as neural network model)
- More detailed, more targeted dataset can be used for a specific region. Even though the dataset we used was provided by World Bank with over 1000 indicators, the spread of data entries is not equal across all nations.

Conclusions

It is possible to predict the growth of a nation's cellphone market by looking at other economic/social indicators

From global perspective, highly correlated indicators do not necessary come from any technology industry category.

The top 5 markets for cellphone are: China, India, United States, Russian Federation, Brazil

Acknowledgements

This work was done by Tam N. Nguyen and was :

- Based on the World Development Indicator dataset published by World Bank and reformatted for Jupyter Notebook by Kaggle
- Implemented using Jupyter Notebook with Python 3
- Under the scope of UCSD's "Python for Data Science" course (UCSanDiegoX: DSE200x)

Nguyen's LinkedIn profile: https://www.linkedin.com/in/tamcs/

References

Jupyter Notebook source codes:
https://github.com/genterist/DataScience/blob/master/CellphoneMarkets/Cellphone_Market_Research.ipynb

World Bank’s World Development Indicators:
https://data.worldbank.org/data-catalog/world-development-indicators

World Bank’s category codes:
http://databank.worldbank.org/data/download/site-content/WDI_CETS.xls

Jupyter Notebook installation guide:
http://jupyter.org/install.html